|CURIEUX|

ACADEMIC JOURNAL

April Issue

Part 2 Issue 49

Editing Staff

Chief-In-Editor

Caroline Xue

George Hasnah

Chief of Operations

Anshal Vyas

Assisting Editors

Olivia Li

Luzia Thomas

Madalyn Ramirez

Shefali Awasthi

Soumya Rai

Sarah Strick

Table Of Contents

The Role of Transportation Policy in Reducing Carbon Emissions: An Evaluation of Effectiveness By Danielle Chen

Abstract

Transportation, a cornerstone of the global economy and the second most significant contributor to global carbon emissions — at more than one third of total emissions — plays a sizable role in climate change. While many states have developed transportation policies to mitigate carbon emissions, the effectiveness of these policies continues to be up for debate. Through a comparative analysis method, this paper explores transportation policies from the European Union (France) and China as well as the impact of these policies on carbon emission reduction. Subsidy-based EV policies are more effective in countries with coordinated infrastructure investments.

Keywords: Transport policy, carbon emissions, policy effectiveness, China, EU, France.

Introduction

The role of carbon pollution in raising the planet's temperature has become a pressing concern in recent years. Burning fossil fuels like coal, oil, and gas is one of the main drivers of climate change, responsible for over 75% of all greenhouse gas (GHG) emissions and nearly 90% of carbon dioxide pollution (United Nations, n.d.). These emissions are fueling increasingly severe weather events. For instance, heavy rains in China caused devastating floods that destroyed towns and cities in 2020, Australia endured catastrophic wildfires that burned millions of hectares of land in 2019 and 2020, and deadly heatwaves in India and Pakistan claimed thousands of lives in 2022.

Beyond environmental damage, climate change now poses a serious threat to global safety. As disasters become more frequent, challenges such as water shortages and shrinking farmland are intensifying, making life harder for many people and increasing the risk of conflict between nations. Reducing energy consumption and cutting carbon pollution is no longer a choice—it's an urgent necessity that demands serious and sustained action.

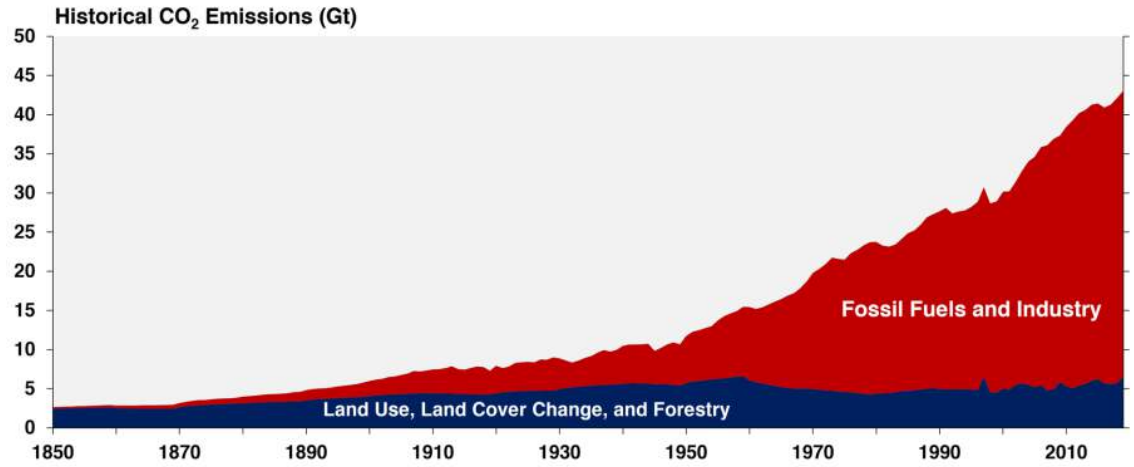


Figure 1.Source: Data from IPCC (2022); Based on global emissions from 2019, details on the sectors and individual contributing sources can be found in the Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, Mitigation of Climate Change, Chapter 2.

As a critical sector in global climate governance, the transport industry's carbon lock-in effect has become a systemic challenge. According to the latest data from the U.S. Environmental Protection Agency (EPA, 2024), 94.3% of the sector's energy supply still relies on gasoline and diesel combustion, contributing 24.7% of global direct greenhouse gas emissions. When accounting for indirect emissions linked to electricity use, its overall share rises to the second-highest among all sectors.

This structural dependence is deeply rooted in historical development patterns. As illustrated in Figure 1, data from the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC, 2022) show that cumulative emissions from fossil fuels and industry have surged sharply since 1950. This turning point aligns with the exponential expansion of global transport infrastructure: between 1950 and 2000, the length of paved roads increased by 412%, air passenger traffic grew 58-fold, and containerization in maritime shipping surpassed 60% (World Bank, 2023).

The entrenched reliance on internal combustion engine technology has not only locked in the sector's emission profile but has also intensified urban air pollution through co-emitted pollutants such as PM2.5 and nitrogen oxides, creating a dual burden on climate and public health. At present, emissions from this sector continue to rise at an average annual rate of 1.7% (IEA, 2023), with its slow decarbonization progress representing a major gap in meeting the temperature control targets of the Paris Agreement.

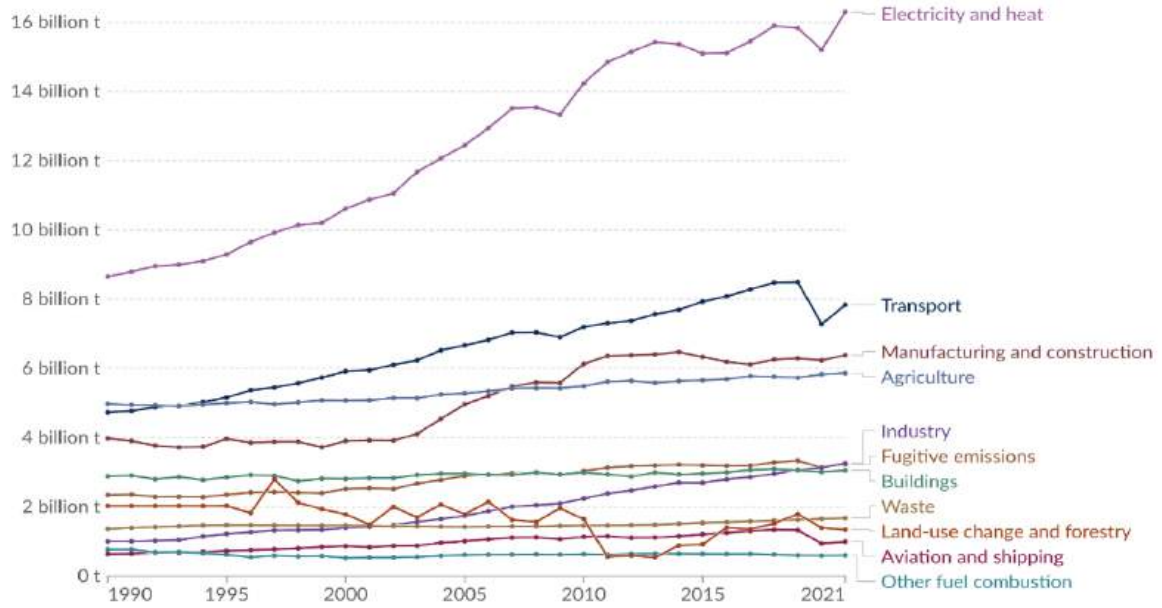


Figure 2.Climate Watch (2024) – with major processing by Our World in Data. "Agriculture" [dataset]. Climate Watch, "Greenhouse gas emissions by sector" [original data].

Electricity and heat sector remains the largest source of emissions, with a consistent upward trajectory from 1990 to 2021(figure 2). This category includes cars, motorcycles, and buses, which together make up 60% of road transport emissions.

Consequently, governments and organizations worldwide have implemented a variety of measures to reduce carbon emissions. These include setting emission caps, managing carbon quotas, applying carbon taxes, and creating carbon trading systems.(Parry, I.W.H., Black, S., & Zhunussova, K. 2022). These measures include emission caps, carbon quota management, carbon taxation, and carbon trading systems.

- **Emission Caps**
  Governments set overall limits on carbon emissions for specific industries or entire economies. Over time, these caps are tightened, compelling businesses to adopt cleaner technologies. For instance, the European Union mandates that member states reduce carbon emissions from the transport sector by 45% by 2030 compared to 2005 levels. Similarly, France's *Climate and Resilience Law* requires a 40% reduction in transport-related emissions by 2030 relative to 2015.
- **Carbon Quotas**
  This system allocates a total emissions limit into tradable allowances, requiring companies to hold permits corresponding to their emissions. China launched its national carbon market in 2021, initially covering 2,162 power sector enterprises with an annual quota of approximately 4.5 billion tons of CO_2. The market mechanism determines carbon pricing, with the average price reaching around 60 yuan per ton in 2023.
- **Carbon Tax**
  A tax is imposed on fossil fuels based on their carbon content. Sweden introduced a carbon tax in 1991, which by 2023 had reached 1,300 Swedish kronor per ton of CO_2 (approximately $120). This policy has contributed to a 35% reduction in oil consumption in the country's transport sector since 1990.
- **Carbon Trading Systems (ETS)**
  The European Union Emissions Trading System (EU ETS) extends to the aviation sector, requiring airlines to purchase permits covering 100% of emissions from intra-EU flights starting in 2024. In 2023, carbon allowance auctions for the aviation industry generated €1.6 billion in revenue, of which €1.2 billion was allocated to research and development of sustainable aviation fuels (European Commission, 2023).

Although many countries have implemented policies like carbon emission limits and carbon taxes to reduce pollution, there is little research on how effective these measures are in the transportation sector, especially when comparing different countries. My study aims to address this issue by first analyzing the main sources of transportation-related carbon emissions, such as the proportion contributed by cars, trucks, and other land-based vehicles. Then, I will

compare the transportation environmental policies of China and France (representing the EU) to determine which measures are truly effective and which may need further improvement.

Methods

This research analyzes quantitative data from the International Energy Agency (201EA) and national transport ministries (2015-2023) with qualitative evaluation of France's and China's EV incentive programs. Through reviewing peer-reviewed articles, policy documents (e.g., EU Fit for 55 package, China's 14th Five-Year Plan), and emission datasets, the research will focus on these following aspects: 1) emission reduction rates, 2) EV adoption growth, 3) charging infrastructure density, 4) policy cost per ton of CO_2 reduced, and 5) social equity impacts. France was selected as a developed economy with carbon taxation leadership, while China represents developing nations combining manufacturing scale and aggressive infrastructure investments, enabling cross-context effectiveness testing of subsidy-based decarbonization strategies.
This study addresses two key questions:

1. How do France's and China's EV subsidy policies differ in design and implementation?
2. Which approach achieves greater carbon reduction per dollar spent?

We hypothesize that France's policy combining subsidies with charging infrastructure reduces emissions more efficiently than China's subsidy-only model.

1.Structure of traffic emissions and pathways for the energy transition

Global transportation-related carbon emissions exhibit significant structural differences across modes. Although air travel accounts for only 2–3% of total passenger transport, it contributes 12% of the sector's emissions—each transatlantic flight passenger generates over one ton of CO_2, nearly three times the annual carbon footprint of an average Indian resident (ICAO, 2023). Ground transportation, due to its sheer scale, remains the dominant source, with fuel-powered vehicles alone emitting 6.5 billion tons of CO_2 annually, representing 74% of the sector's total emissions (IEA, 2023).

This emission pattern is deeply rooted in long-standing reliance on fossil fuels. Currently, 94% of transportation still depends on gasoline and diesel, despite the fact that diesel engine emissions contain PM2.5 particulates, classified by the WHO as a Group 1 carcinogen. Additionally, refrigerants used in vehicle air conditioning contribute to 7% of ozone-depleting substances released into the atmosphere (WHO, 2022; UNEP, 2021).

However, energy diversification is beginning to disrupt this dependence. While alternatives such as biofuels and hydrogen are gradually entering the market, the real transformation is being driven by the rapid rise of **electric vehicles (EVs)**. In 2023, one in every five new cars sold in the EU was rechargeable, with battery electric vehicles (BEVs) surpassing a 14.6% market share. This shift is not solely the result of technological advancements but also reflects strategic policy design. As China's large-scale battery production has reduced EV costs

by 89% over the past decade, and France has leveraged its nuclear energy capacity to keep the lifecycle emissions of EVs at just 15% of those from gasoline-powered cars, the balance of the market is beginning to tilt (CATARC, 2023; RTE, 2023).

2**. Why choose electric vehicles (EVs)?(what are electric vehicles (EVs))**

Electric Vehicles (EVs) are transport modes powered by battery energy and electric motors, producing zero direct emissions during operation. By 2023, electric vehicles (EVs) had emerged as the leading replacement for gasoline and diesel cars, representing more than 14.6% of new car sales in the European Union. As a result, over 20% of new cars sold in Europe were chargeable.(figure 4)

Electric vehicles (EVs) are mainly divided into three categories: traditional hybrid vehicles, plug-in hybrid electric vehicles (PHEVs), and battery electric vehicles (BEVs). Each type has its own unique features and functions.

- Traditional Hybrid Vehicles: These cars, often just called hybrids, combine an internal combustion engine with an onboard electric motor. The electric motor can convert kinetic energy into electricity during braking and store it in the battery. Compared to regular gasoline-powered cars, hybrids improve fuel efficiency by about 40%. However, they still rely primarily on gasoline or diesel and cannot be charged from an external power source.
- Plug-in Hybrid Electric Vehicles (PHEVs): PHEVs have larger batteries than traditional hybrids and can be charged using an external power source. They can run purely on electric power for 20 to 50 miles before switching to a gasoline or diesel engine as a backup. In hybrid mode, their fuel efficiency is similar to that of traditional hybrid vehicles.
- Battery Electric Vehicles (BEVs): BEVs run entirely on electricity stored in rechargeable battery packs. They don't have an internal combustion engine and produce no tailpipe emissions.

EVs have great potential to reduce carbon emissions, especially when powered by cleaner energy sources. EVs produce zero tailpipe emissions when running entirely on electricity, while gasoline cars emit over 15,000 pounds of carbon dioxide per year, making them the highest emitters among all vehicle types.

The indirect emissions from EVs depend on the energy sources used in the power grid. 38.32% of electricity in the grid comes from natural gas, and 19.85% from coal—both major sources of carbon emissions. However, renewable energy sources like wind (10.47%) and solar (4.90%) help reduce the impact of these emissions. In contrast, gasoline-powered cars rely entirely on fossil fuels, keeping their emissions consistently high.

EVs play a central role in global transportation decarbonization policies due to three interconnected factors. First, their direct technological substitution potential—according to the International Energy Agency (IEA, 2023), electrification of light-duty vehicles could contribute

45% of the emissions reductions required in the transport sector by 2030. Second, the efficiency of policy intervention—as seen in the European Union's carbon trading system, which has made the carbon costs of fuel-powered vehicles explicit, leading to EVs achieving a lower total lifecycle cost than conventional vehicles for the first time in 2022 (ICCT, 2022). Third, the industrial synergy leverage—China's strategic expansion of battery production capacity has led to an 89% reduction in EV manufacturing costs over the past decade (CATARC, 2023), accelerating market penetration.

The focus on EV policy evaluation is not only due to their high technological maturity but also because they simultaneously offer verifiable carbon reduction metrics—such as the EU's Well-to-Wheel standard—and controllability in supply chain restructuring, with China holding 67% of global patents on rare-earth permanent magnet motors. These factors position EVs as a critical case study for addressing the challenge of delayed adoption of **l**ow-carbon transport technologies. The feasibility of this transformation is concentrated in the field of electric vehicles: while France relies on nuclear power to reduce the full-cycle emissions of EVs to 15% of those of fuel vehicles (RTE, 2023), China has achieved an 89% reduction in battery cost per kWh through economies of scale (CATARC, 2023). The synergistic mechanism of technological innovation and policy tools is already clear.

3. About transportation emission reduction policies

As a key way to reduce emissions from transportation, electric vehicles (EVs) have become an important strategy for addressing climate change. This is due to their potential to cut greenhouse gas emissions and improve air quality. So far, the main policies driving EV adoption have been financial incentives, which have not only supported the early rollout of electric light-duty vehicles (LDVs) but also encouraged the growth of the EV manufacturing and battery industries.

These financial incentives are designed to narrow the price gap between EVs and traditional gasoline cars. Common examples include purchase subsidies and tax breaks on buying or registering vehicles. For instance, Norway introduced such measures as early as the 1990s, while the U.S. started in 2008, and China followed in 2014.

To accelerate the low-carbon transition in transportation, France and China have developed distinct electric vehicle (EV) subsidy systems. France's "Bonus écologique" policy is based on fuel tax redistribution, using revenue from gasoline taxes to fund EV purchase subsidies. It also includes emission standards and social fairness measures—high-income groups receive only the basic subsidy, while low-income households qualify for additional support.

China, on the other hand, has followed a phased industry development strategy. In the early stage, high government subsidies helped rapidly expand the market. In the mid-term, technical standards and gradual subsidy reductions guided industrial upgrades. Finally, the system shifted to market-based regulation, such as carbon credit trading.

France focuses on a "polluter pays" model using tax incentives for a fair transition, while China prioritizes policy adjustments to build a complete EV industry chain. These differences make France and China valuable case studies for comparative research.

- **Bonus écologique (Ecological Bonus**）

France's Bonus écologique is one of the most comprehensive electric vehicle (EV) incentive programs in Europe, focusing on strict technical standards and regional fairness. According to the 2023 update by France's Ministry of Ecological Transition, fully electric vehicles (BEVs) must meet a strict CO_2 emission standard of ≤20g/km (equivalent to an energy consumption of ≤15kWh/100km under WLTP testing) to qualify for subsidies. Plug-in hybrid vehicles (PHEVs) are allowed a higher limit of ≤50g/km to receive a basic subsidy of €1,000. This tiered standard encourages consumers to choose zero-emission technologies.
Between January 2021 and December 2022, the program also provided up to €50,000 in special subsidies for electric trucks over 3.5 tons, leading to a 217% increase in heavy-duty EV registrations within two years. However, as of 2023, the focus shifted to light passenger vehicles, and the subsidy for heavy-duty trucks was discontinued. Additionally, residents in overseas regions, such as Réunion, are eligible for an extra €1,000 subsidy, which will remain in place until 2025. This policy contributed to Réunion's EV penetration rate reaching 82% of mainland France's level by 2023.

In terms of funding, France adopted the "polluter pays" principle. Each year, €1.47 billion is allocated to the subsidy pool from the fuel tax (TICPE), meaning that every gasoline car owner indirectly contributes €47 annually to the development of EVs. This "fuel-to-electricity" funding cycle has helped France's BEV market share grow steadily from 1.2% in 2017, when the policy was introduced, to 16.8% in 2023. At the same time, public charging infrastructure expanded to 12.3 charging points per 100 kilometers of road, creating a positive feedback loop between infrastructure and market demand.

- **Promotion of new energy vehicles (china)**

China's strategy for promoting new energy vehicles (NEVs) reflects a distinct policy evolution shaped by the realities of a developing economy. The initiative began in 2009 with the "Ten Cities, Thousand Vehicles" pilot program, which focused on public transportation as an entry point. Under this scheme, the government provided a ¥500,000 subsidy per electric bus, covering approximately 40% of the vehicle's cost at the time.
In 2014, the policy shifted toward the private market, introducing a minimum driving range of 150 km as a qualification for subsidies, which ranged from ¥35,000 to ¥60,000 per vehicle. This directly led to the rise of mass-market electric models like the BAIC EU series. However, after a 2016 subsidy fraud scandal, the policy framework adopted dynamic adjustments:

- From 2017 to 2019, the range requirement increased by 20% annually (from 150 km to 250 km), while subsidies were gradually reduced.

- By 2020, a precise subsidy phase-out plan was introduced—private car purchase incentives declined 30% over three years, while public transport vehicles retained a maximum ¥500,000 subsidy to ensure that over 90% of buses were electrified by 2022.

In 2023, China marked a historic policy shift. After investing over ¥300 billion in subsidies, the nationwide purchase subsidy program officially ended, transitioning instead to a market-driven approach through the "Dual Credit" system. Under this policy, automakers were required to increase their NEV credit ratio from 14% in 2021 to 18% in 2023, effectively forcing traditional car manufacturers to pay around ¥3,000 per fuel-powered vehicle in NEV compliance credits.

The fundamental shift behind this transition was the maturity of the industry. Battery costs fell from ¥3.5/Wh in 2014 to ¥0.85/Wh in 2023, enabling NEV adoption to continue growing at 35% annually even after subsidies ended.

One of the most innovative policy moves was favoring battery-swapping technology. While most EV subsidies were capped at ¥300,000 per vehicle, swappable-battery models like the NIO ES8 were exempt from this limit, provided that their battery assets were managed separately. This “vehicle-battery separation” model helped increase the daily operating mileage of electric taxis by 40%, proving to be a viable technological breakthrough in the post-subsidy era.

Policy analysis

- **Bonus écologique (Ecological Bonus**）

Year	BEV Sales Share	Subsidy Adjustment	Annual Sales Growth Rate
2019	2.80%	€ 6,000	38%
2020	6.70%	€ 7,000	140%
2021	9.80%	€ 6,000	46%
2022	13.30%	€ 5,000	36%
2023	16.80%	€ 5,000	26%

figure3. French Automobile Industry Association (PFA, 2024)- (The subsidy in 2020 was temporarily increased due to the pandemic.

For every additional €1,000 in subsidies, the BEV market share increased by approximately 3 percentage points (elasticity coefficient 0.3, 2019–2023).
During the heavy-duty vehicle subsidy period (2021–2022), electric truck sales grew by 217%. (figure 3)

- **Promotion of new energy vehicles (china)**

Year	NEV Penetration Rate	Annual Sales (10,000 units)
2014	0.30%	7.5
2016	1.80%	50.7
2019	4.70%	120.6
2021	13.40%	352.1
2023	31.60%	949.5

figure 4.China Association of Automobile Manufacturers (CAAM, 2024)
The cost of CO_2 reduction per ton dropped from ¥8,000 in 2014 to ¥3,200 in 2022, driven by economies of scale.

In 2023, with the removal of subsidies, the market continued to grow, with EV sales increasing by 35.7%, indicating strong independent market demand.(figure4)

Result

France and China have taken very different policy paths, leading to distinct results. France followed a "regulation + infrastructure-first" strategy, using a dense charging network and flexible subsidies to grow the pure electric vehicle (BEV) market fivefold in five years. One of its most innovative policies, the “fuel tax redistribution” system, redirected gasoline tax revenue to fund EV subsidies. However, when subsidies for heavy-duty EVs were removed, sales dropped sharply, revealing the limitations of relying on a single policy tool.
China, on the other hand, used a "industry development + market adaptation" strategy. With battery costs falling by nearly 90%, China pushed NEV adoption from almost zero to over 30%, even after subsidies ended. This proves that well-designed policies can create lasting market momentum.

Efficiency and Fairness

When comparing policies, France and China show an opposite balance between efficiency and fairness. France, with nuclear energy as its main power source, has a much lower carbon reduction cost—only one-third of China’s. However, low-income households own far fewer EVs, making social fairness a weakness of the policy.
China improved fairness through programs like “EVs for Rural Areas”, raising EV adoption in smaller cities to 60% of first-tier city levels. But because China’s electricity still relies heavily on coal, EV lifecycle emissions are seven times higher than in France, reducing overall climate benefits. This trade-off highlights that policies must be designed to fit local energy resources and social conditions.

the effects After Subsidy Removal

The way each country handled the end of subsidies also reflects different governance styles. France passed laws to phase out 20% of high-emission vehicles, but this slowed market growth by half. China, instead, built a carbon credit trading system worth over ¥30 billion per

year, pushing traditional automakers to transition, which kept EV sales growing even after subsidies ended.

Discussion

- .Reassessing the Environmental Benefits of EVs

Although EV production generates 15–30% more carbon emissions than conventional fuel-powered vehicles, their lifecycle emissions remain significantly lower. In the U.S. energy mix, EVs can reduce emissions by 60–68%, with smart charging strategies (e.g., off-peak charging) further improving efficiency by 18%. France's nuclear-powered grid allows EVs to achieve an 85% emission reduction, while China's reliance on coal power weakens this advantage. However, advancements in battery technology—such as CARL's Qilin battery, which increases energy density by 13%—and improvements in recycling systems—with lithium recovery rates rising from 5% to 35% by 2023—are gradually closing the gap. These trends indicate that the environmental benefits of EVs depend on a dual strategy of clean electricity and technological innovation.

- Research Limitations and Future Directions

This study has three key limitations:

1. Exclusion of the aviation sector, which contributes 12% of transportation-related emissions;
2. Lack of quantification of China's battery exports, which may generate global emission reduction spillover effects;
3. Short policy assessment timeframe, limited to 2023, preventing analysis of long-term breakthroughs such as solid-state batteries.
4.

Future research could expand into comparative studies of aviation decarbonization policies and incorporate Dynamic Lifecycle Assessment (DLCA) to track the global diffusion of battery innovations.

Conclusion

This research examines the impact of transportation policies on carbon reduction, focusing on France and China's electric vehicle (EV) promotion strategies as representative cases. The findings show that while both countries have made significant progress in decarbonizing transport, their approaches differ in policy design, efficiency, and social equity. France's Bonus écologique policy follows a "regulation + infrastructure-first" approach, combining strict technical standards, fuel tax redistribution, and charging network expansion. Between 2017 and 2023, these measures helped increase the market share of battery electric vehicles (BEVs) from 1.2% to 16.8%—a fivefold growth. Thanks to its nuclear energy dominance, France has achieved an 85% reduction in EV lifecycle emissions, with a cost per ton

of CO_2 reduced (static efficiency) at just €230. However, the policy has equity shortcomings, as EV adoption rates among low-income households remain significantly below the national average, revealing the challenge of balancing environmental and social priorities.

China, in contrast, adopted a phased subsidy strategy that prioritized industrial scaling and market adaptation. By combining generous government subsidies, incentives for battery-swapping technology, and a shift toward carbon credit trading, China managed to cut EV manufacturing costs by 89% (2014–2023) and sustain a 35.7% year-on-year sales increase even after subsidy removal. While the cost per ton of CO_2 reduction (dynamic efficiency) decreased by 60% (to ¥3,200), China's coal-heavy electricity grid weakened the climate benefits, with EV lifecycle emissions seven times higher than in France.

The comparison highlights an important takeaway: policy design must align with a country's energy resources and socioeconomic conditions. France's regulatory precision offers a model for nations with abundant clean energy, whereas China's large-scale industrial strategy demonstrates that developing countries can break technological lock-in through government intervention—though this comes with environmental trade-offs during the transition.
The effectiveness of EV policies depends on a dual approach: accelerating the transition to cleaner electricity sources to maximize carbon reductions while ensuring that social equity measures make EV adoption accessible to a broader population. As countries refine their climate strategies, the French and Chinese experiences highlight the importance of flexible, locally tailored policies in advancing sustainable transportation.

Works Cited

Notice on the Issuance of the Technical Policy on the Prevention and Control of Motor Vehicle Emission Pollution (Huan Fa [1999] No. 134, 2003),

United Nations. (n.d.). Climate Change. Retrieved from https://www.un.org/climatechange

Intergovernmental Panel on Climate Change (IPCC). (2022). Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, Mitigation of Climate Change, Chapter 2. IPCC.

U.S. Environmental Protection Agency (EPA). (2024). Report on Greenhouse Gas Emissions. Retrieved from https://www.epa.gov/ghgemissions

Ritchie, H., Rosado, P., & Roser, M. (2020). Emissions by Sector. Our World in Data. Retrieved from https://ourworldindata.org/emissions-by-sector

Climate Watch. (2024). Greenhouse gas emissions by sector. Retrieved from https://www.climatewatchdata.org/ghg-emissions

Parry, I.W.H., Black, S., & Zhunussova, K. (2022). Economic Costs of Carbon Emissions. Journal of Environmental Economics and Management.

Huan Fa. (2003). Air Pollution and Health Risks from Vehicles. China Environmental Science Press.

Ruixie. (2017). Transport Infrastructure and Economic Development. Journal of Transport Economics and Policy.

Ritchie, H. (2023). Aviation and Carbon Emissions. Environmental Research Letters.

International Energy Agency (IEA). (2023). Global EV Outlook. Retrieved from https://www.iea.org/reports/global-ev-outlook-2023

International Council on Clean Transportation (ICCT). (2022). Electric Vehicles: Costs and Benefits. Retrieved from https://theicct.org/electric-vehicles-costs-and-benefits-2022

China Automotive Technology and Research Center (CATARC). (2023). Electric Vehicle Manufacturing Costs. CATARC Reports.

Alternative Fuels Data Center. (2023). Electric Vehicle Emissions. Retrieved from https://afdc.energy.gov/vehicles/electric-emissions

Moseman, A. (2022). Electric Vehicles and the Environment. Clean Tech Magazine.

Federal Highway Administration (FHWA). (n.d.). Transportation and Climate Change. Retrieved from https://www.fhwa.dot.gov/policy/otps/innovation/issue1/policies.cfm

Energy and Clean Air Solutions (ECAS). (2021). The French Recovery Plan: Transport Sector. Retrieved from https://clean-energy-islands.ec.europa.eu/countries/france/legal/res-transport/subsidy-bonus-ecologique

Air Treatment with CO_2 Capture Methods: an Update of Trends in 2020-2024

By Teenueng Limpanapa

Abstract

Carbon dioxide is a greenhouse gas that plays a significant role in causing global warming and climate change. Accumulation of excess CO_2 can cause future harmful effects; therefore, a method of mitigating climate change, such as decarbonization, is necessary. This review analyzes research efforts made on carbon capture from 2020 to 2024 to compare existing methods and identify the most efficient and effective one. The rise in CO_2 emissions is strongly associated with the constant demand for fossil fuel usage globally over the years. Emphasis is placed on significant trends and innovations in Direct Air Capture (DAC). Additionally, trends of Point-Source Capture techniques that trap CO_2 straight from an emission source are analyzed.

Keywords

Earth and Environmental Sciences; Climate Science; Carbon Capture; Point Source Carbon Capture; Direct Air Capture

Introduction

Carbon dioxide (CO_2) is a colorless, non-flammable, and acidic gas, covalently double bonded by Carbon and Oxygen.[1] Accumulating in the atmosphere, CO_2 traps heat that reflects from the surface inside Earth's atmosphere to maintain a habitable temperature for humanity (Figure 1). The constant increase in greenhouse gas emissions leads to an increase in Earth's average temperature, referred to as global warming.[2] According to the Environmental Impact Assessment (EIA), the current CO_2 level of 425 ppm exceeds the preset levels of atmospheric CO_2 by approximately 100 ppm.[3] The American Conference of Governmental Industrial Hygienists (ACGIH) advises an 8-hour Time-Weighted Average (TWA) Threshold Limit Value (TLV) of 5,000 ppm, with a peak exposure limit of 30,000 ppm for up to 10 minutes.[4] A concentration of 40,000 ppm is deemed life-threatening and hazardous to health (IDLH level). Furthermore, it is suggested that CO_2 levels in buildings can reach up to 1,000 ppm but must not exceed 1,500 ppm, as it can cause danger to humans.[5] Thus, decarbonization, as an effort to reduce CO_2 emissions, is a significant concept.

The concept of decarbonization was first introduced in the 2015 Paris Climate Agreement, as governments were "pursuing efforts to limit temperature increase to 1.5°C above pre-industrial levels."[6] Decarbonization is defined as the process of reducing or eliminating CO_2 emissions. Increased efforts are dedicated to the research of decarbonization technologies, such as Direct Air Capture (DAC) and Point-Source Carbon Capture, during 2020-2024. The concept's implications are used in hopes of attempting to reduce the amount of CO_2 in the air, creating a balance between gases to improve the surrounding air.

A diverse collection of carbon capture methods has been introduced, with the most notable being Direct Air Capture and Point Source Carbon Capture.[7] The major difference in the

classification of CCS projects is the source of CO_2, where Point Source Carbon Capture extracts carbon from an emission source, like chemical plants or coal power plants, while DAC extracts carbon from the ambient air.[8] The current DAC technology is being developed under the innovation of new basic sorbents used to extract CO_2. Some of the latest innovations involve the use of solid-supported amines, metal-organic frameworks (MOFs), and alkali, as well as alkaline-earth compounds like $Ca(OH)_2$, KOH, and NaOH.[9] In contrast, the differentiations between post-, pre-, and oxyfuel combustions are not as strongly researched as Direct Air Capture methods due to their reliability on an emission source. Oxyfuel and post-combustion capture methods are interchangeably similar, such that the fuel is combusted with nearly pure oxygen rather than atmospheric air during post-combustion. Pre-combustion capture occurs when the fuel is converted into a synthetic gas consisting of CO, CO_2, and H_2 under the pressure of steam.[10]

A newly proposed DAC process utilizes calcium hydroxide ($Ca(OH)_2$) and calcium carbonate ($CaCO_3$) to capture CO_2 from the atmosphere. Ambient air will circulate through porous structures composed of $Ca(OH)_2$, enabling CO_2 to gradually react and form $CaCO_3$.[11] Additionally, CO_2 post-combustion capture is proposed using methods such as chemical absorption, physical absorption, and membrane separation, whilst pre-combustion capture includes water-gas shift followed by CO_2 extraction from the gas.[12] A majority of CO_2 capture methods often integrate the mechanisms of absorption and chemical reactions in order to extract CO_2 from the atmosphere.

In previous years, researchers have been introduced to Carbon Capture and Storage (CCS). This refers to a set of technologies that focuses on the selective removal of CO_2 from gas streams, its compression into a super-critical condition, and finally, its transportation and sequestration in geologic formations.[13] The current trend now shifts towards Carbon Capture and Utilizations (CCU), which focuses on extracting CO_2 and then transporting it through a pipeline for utilization instead of storing it so that it can't be released back into the atmosphere like CCS.[14] The utilization of CO_2 is seen as an opportunity to create economically viable and sustainable solutions, in addition to introducing a cost-effective pathway of both capturing and utilizing the greenhouse gas. This allows the captured gas to be converted into synthetic fuels, provides a cleaner alternative to traditional fossil fuels, manufactures carbon-based materials, and facilitates the chemical synthesis of valuable chemicals and pharmaceuticals. Although it is an upcoming research trend, carbon utilization is not the main focus of this paper.[15]

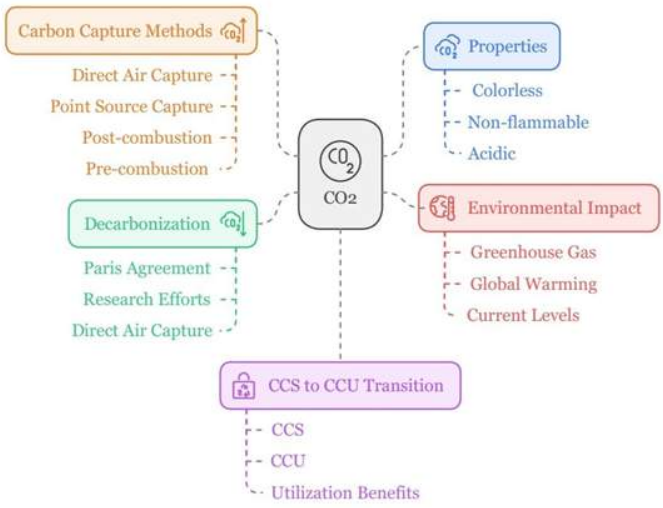


Figure 1: Summary of Carbon Dioxide and its impacts.

Discussion

1. Historic Decarbonization Methods

The Energy Information Administration suggests that there are four pillars of decarbonization: energy efficiency, industrial electrification, low-carbon fuels, feedstocks and energy sources, and carbon capture, utilization, and storage.[16,17] Energy efficiency is considered the most cost-effective option for reductions in greenhouse gas emissions. However, this is a seemingly difficult concept since it involves behavioral changes of consumers. For instance, transitioning to more efficient public transportation options would account for most of the behavioral changes needed to achieve net zero emissions by 2050.[18] Furthermore, Industrial electrification involves the substitution of fossil fuel-based technology or processes with electrically powered equivalents on an industrial scale. This would lead to large-scale increases in renewable electricity or nuclear power supply and a new infrastructure for the world. Low-carbon fuels, feedstocks, and energy sources can lead to reduced combustion-associated emissions for industrial processes, which introduces new research on the use of biofuels and bio feedstocks. Lastly, Carbon Capture, Utilization, and Storage (CCUS) describes a multi-faceted approach to extracting CO_2 and using the captured carbon to generate value-added products, known as utilization, or storing it for an extended period.

1.1 BECCS Technology

When CCS is combined with neutral bioenergy (BECCS), it can generate negative emissions. Figure 2 depicts the challenges BECCS technology could overcome: its ability to reduce emissions and produce energy makes it an efficient choice for implementation. The procedure involves extracting carbon from ambient air, transferring it into a biosphere, and storing it permanently in a geosphere. Using the model of intercomparison study by Koelbl et al., certain general statements on the usage of CCS in integrated assessment models (IAMS) can be

made. This model suggests that the amount of CO_2 sequestered by biomass depends on the applied capture rate, the content of biomass carbon content, the heating value of carbon content, and the biomass carbon footprint. Its high share in portfolios with low stabilization would likely contribute to less deployment of BECCS under the assumption that there are equally cost-effective alternatives.[19]

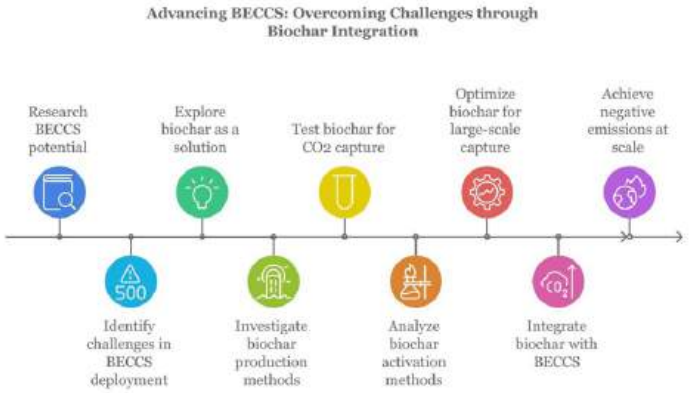


Figure 2: Advancing Bioenergy with Carbon Capture and Storage to overcome challenges.

In the past years, BECCS has been widely researched. The permanent capture and sequester of CO_2 are claimed to create negative emissions, as stated by Freer et al. However, it is notable that no two facilities are the same, and there are multiple key factors affecting its performance.[20] Generic success of a BECCS supply chain in a specific location is determined by encompassing local infrastructure, accessible biomass, economic conditions, energy systems, land management approaches, socio-political factors, and geographical aspects. Although the extraction facilities remain on land, the captured carbon is transported to an offshore geological storage facility. This implies that there is a limited number of areas in which BECCS can be deployed. From a spatial context, a 10 km relocation of the facility leads to an increase in spatially explicit supply chain emissions by 9-13%. Furthermore, the BECCS facilities with low purity CO_2 at high yields result in reduced spatial emissions, while high purity CO_2 at low yields is produced at higher spatial emissions.

From this research and analysis, BECCS encompasses multiple difficulties, from deploying CCS technology in a biomass facility to finding a suitable site for sufficient yield extraction. Even though there is high encouragement from multiple nations due to the towering capture scale at 1 $MtCO_2$ per yr (Mt = million tons), there remains an insufficient number of plans for deployment across all sectors. The two main constraints are economic viability and land use competition. Firstly, CCS technology alone is already costly; additional costs in spatial contexts, along with transportation and storage, may not be as financially competitive as other carbon reduction methods. Moreover, the technical and geographical limitations on siting create a restriction on which locations are feasibly available. These are the reasons that are assumed to indicate why BECCS is not implemented globally.

There is a myriad of disadvantages associated with the large-scale deployment of BECCS. As a result, extensive research is being undertaken to investigate the integration of biomaterials with CCS. For instance, scientists have discovered biochar that is manufactured through pyrolysis. Pyrolysis is a thermochemical conversion of biomass, strongly impacting the resultant characteristics of its product. The process is categorized into slow and fast, referring to the rate at which heat is applied: the natural polymers in biomass will transform, break apart, and fragment at varied temperatures.[21] This serves as an alternative, as researchers are able to transform waste into valuable products. Through the literature review, the yield of biochar synthesis is relatively low, with only high yields produced in slow pyrolysis.[22] Zhang et al.'s research tested the capture capacities of biochar through ASAP 200, a Kane 457 gas analyzer, and a thermal gravity device for three categories of biochar. The resulting graph illustrates a consistent trend in terms of adsorption, presenting the effect of different activation methods. In addition, the adsorption capacity of wood pellet biochar is observed to increase by 86% when activated using water vapor. The paper suggested that the varied contextual testing may have differing results due to pore collapse during activation or an increase in surface acidity. Despite activation enhancing pore structures, increasing capture capacity on a larger scale is subjected to other external factors, such as surface functionality.[23]

1.2 Point Source Capture

Point Source Capture is the extraction of CO_2 from a specific emission source. Figure 3 illustrates the breakdown of the point source capture method into pre-combustion, post-combustion, and oxyfuel combustion capture. These are categorized depending on where CO_2 capture occurs. Pre-combustion capture occurs before combustion, so the carbon is removed physiochemically. Alternatively, post-combustion is where CO_2 is captured after fuel combustion. Lastly, oxyfuel combustion is interchangeably like post-combustion capture, but the fuel is combusted with pure oxygen. This results in a cleaner captured emission. An advantage of point source capture is that it can capture over 40 $MtCO_2$ per year on currently operational generation sources, with a possible capture rate of 110 $MtCO_2$ per year in development for future generations.[24]

Pre-combustion capture is used when incomplete combustion occurs in a reactor and transforms into syngas, a synthetic gas composed of CO and H_2. An integrated gasification combined cycle (IGCC) is often used in this approach. Steam and oxygen are provided to the gasifier to produce syngas, where their impurities are later filtered in a cyclone separator. The conversion of syngas and steam into CO_2 and H_2 occurs in a water-gas shift reactor, where the product is purified in a desulfurization unit. Eventually, the converted CO_2 is captured and directed either for storage or utilization. The remaining hydrogen serves as a fuel in the gas turbine. Despite its effective and efficient separation methods, it is difficult to retrofit an existing powerplant with carbon capture technology since they are distinct processes. This could lead to a possible shutdown due to process failure. To capture CO_2 from syngas, effective adsorbents can

achieve more than 90% CO_2 capture whilst reducing plant efficiency. Additionally, high costs are involved due to the installation of a gasification chamber.[25]

In contrast, post-combustion capture extracts CO_2 from flue gas emitted by combustion plants. Flue gas contains the reaction products of fuel and combustion air, mixed with residual substances to be filtered before being released into the atmosphere. Its capture unit is placed after purification systems. Madejski et al.'s paper summarizes seven subcategories of post-combustion capture: absorption solvent-based methods, adsorption-physical separation, membrane separation, chemical looping combustion (CLC), and calcium looping process (CLP), cryogenic method, application of absorption-based post-combustion capture methods and converting CO_2 into value-added chemicals. Each subcategory involves differing methods and materials, making it suitable for a variety of existing plants, given that there is no interaction with power generation except by increased heat and/or power demand.[26] Post-combustion capture is the most frequently considered; however, due to the low partial pressure of CO_2 in flue gas, the driving force for CO_2 also decreases. It is a well-established technology that can be retrofitted for existing power plants, making it a flexible option for manufacturers.[27-29]

Oxyfuel combustion is comparable to post-combustion capture in that CO_2 is collected after the combustion process. However, this process allows fuel to be burnt in almost pure oxygen instead of atmospheric air. Firstly, the air separation unit isolates oxygen from other gases, producing a purity of approximately 95%, which is used in the combustion chamber. The exhaust gas resulting from combustion comprises CO_2 and water vapor. Condensing the water vapor, the temperature used is higher than ambient conditions, leading to separation and compression of CO_2. The process's application is mainly used on a laboratory scale, with possible alterations escalating to boilers and energy systems with gas turbines. There is significant material demand due to elevated conditions, reduction in efficiency, and high capital expense. Energy consumption in CO_2 is lower than that of other point source captures: the most energy-consuming mechanism of this process is the air separation of oxygen for combustion.[30-32]

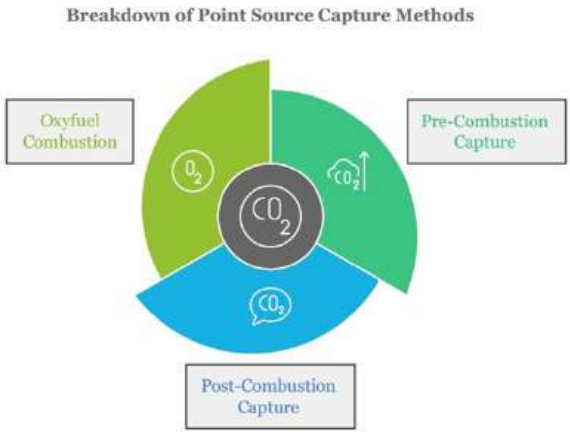


Figure 3: A breakdown of point source capture methods.

2. Decarbonization Trends in 2020-2024

As a recurring global issue, decarbonization is continuously researched through multiple research institutions. Data has shown that there is increasing attention towards the trends of carbon sequestration as a method of storing carbon so that it is no longer released into the atmosphere or can be used for other purposes. The extensive implementation of carbon capture, storage, and utilization technologies has yet to be validated.

2.1 Point Source Capture and Update

The application of physical adsorbents such as zeolite or activated carbon has gained particular attention in recent years. However, these processes are still under research. The adsorption process can either be chemical or physical: both methods require the removal of sulfur dioxide (SO_2) from flue gas cooled to 40-70°C before the carbon adsorption process. The two research fields highlighted are activated carbon, which can be enhanced by adding heteroatoms to boost its CO_2 absorption capacity, and molecular sieves, which separate molecules based on their molecular size. Previous studies evaluated the carbon capture efficiency of steam-activated anthracite, demonstrating that increasing the surface area of anthracite enhanced its low adsorption capacity and chemical modification with ammonia and polyethyleneimine.[33] Conversely, molecular sieves are highly expensive, but they can be tailored to fit nearly any carbon capture process. Particular attention is focused on high-surface-area inorganic support that incorporates basic organic groups, typically amines.[34] Finally, membranes function as semi-permeable barriers that can isolate components from a gas stream through concentration, electric potential, or pressure differences, where kinetic diameters of multiple elements are considered.[35,36] Inorganic membranes are still lab-scale concepts, with porous membranes being commonly applied for carbon capture. Nonporous membranes, however, are utilized for oxygen separation via perovskite systems or hydrogen separation using palladium alloys.[37] Polymeric membranes require less energy than inorganic membranes, with the absence of dangerous emissions and harmful chemicals. They are relevant and suitable for powerplants in post combustion capture and natural gas pre combustion processes.[38,39]

Research in 2020 introduced the idea of carbonation of cement and concrete. Once exposed to the atmosphere, cementitious substances absorb CO_2 via carbonation, although the captured quantity is only a small portion of the gas emitted from calcination during cement manufacture. This process involves calcium-bearing components of cement interacting with CO_2 to form $CaCO_3$ and various non-carbonated compounds. Exposed to the atmosphere, this reaction progressively starts where CO_2 diffuses inwards. Capture-focused studies often assumed a simplified profile of cementitious materials, where some models' maximum carbonation degree varied up to 100%. Lower figures, in comparison, appear more realistic because of elements like porosity and the chemical makeup of hydrates, suggesting the impossibility of reaching such a high percentage. Conceptually, reinventing industrial practices regarding cement and concrete for carbonation to increase CO_2 uptake is possible, as it only lowers the durability of steel when subjected to alternating wet and dry conditions. However, this technique is severely

under-researched, so we are unable to reach a conclusion since multiple factors must be considered before global implementation.[40]

A further literature review by L. Rex et al. in 2021 examined the CO_2 emission reduction technologies, focusing on the iron and steel industry (ISI) of China. The review suggested that post-combustion capture could reduce carbon emissions by an estimated 40%, along with a 20% decrease in the exhaust from the main blower. This contrasts with pre-combustion capture, which only allows for a 20% reduction in carbon emissions. It is predicted that ISI will remain a key area for industrial applications in CCS technology due to its concentrated production model. Graphical models presenting CCS across various industries and sub sectors revealed that ISI could reach up to 10 $GtCO_2$ of carbon capture by 2026. CO_2 accounts for 35% of the total emission in iron and steel plants, implying that pre-combustion capture is a feasible enhancement option.[41,42] However, point source capture's largest disadvantage was its ability to capture only carbon emissions in specific portions and not all emissions generated by the plant.[43]

Compilations of adsorption data for post-combustion capture emphasize the creation and modification of multiple adsorbents. Sorbents are categorized into physisorbents, which consist of carbon-based sorbents, zeolites, metal organic frameworks (MOF), and chemisorbents. In physisorption, molecules of CO_2 are adsorbed on the surface. An example is activated carbon, a solid material noted for its availability, low cost, and renewability but is sensitive to moisture. Zeolites are crystalline microporous substances composed of silicon, aluminum, and oxygen, which have high adsorption capacities and fast kinetics under mild conditions. Lastly, MOFs are crystalline porous substances consisting of organic linkers with metal ions. These are recognized for high tunability and potential for capture. Chemical adsorption of CO_2, however, depends on the nucleophilic nature of functional groups found on surfaces of chemisorbent, especially how they can chemically interact with CO_2. Amine-functionalized adsorbents interact between amine groups and support materials: these are amine-impregnated and amine-grafted sorbents, which enhance CO_2 uptake through chemical interactions. It is concluded that there are still present challenges concerning sorbent materials, the gas-solid interaction system, and regeneration methods.[44]

J.A. Garcia et al.'s paper emphasizes the importance of selecting CO_2 capture technology, which is influenced by several factors. The paper differentiates CO_2 capture into three routes: atmospheric CO_2 capture, precombustion capture, and post-combustion capture. In this context, oxy-fuel combustion is counted as a form of post-combustion capture. Capture technologies considered are chemical and physical adsorption, membranes, and cryogenic separation. The concept of chemical adsorption is based on acid-base neutralization, which is ideally suited for post-combustion systems since there are limitless options for various solvents that have the tendency for absorption, such as MEA and DEA.[45] Physical adsorption does not entail chemical reactions, which makes regeneration of solvents easier than in other carbon capture methods. As a result, it is commercially used to extract CO_2 from syngas and remove acidic gases from natural gas. Pressurized water scrubbing using compressed biogas is a desirable method for pre-combustion capture from biogas. Due to high water requirements, extra wastewater is

produced, which may clog due to bacterial development.[46] Utilizing organic solvents that absorb CO_2, water, sulfur, and aromatic components is an additional option. This process occurs at low temperatures, and case studies do not involve either the Rectisol process (chilled methanol) or the Selexol process (dimethyl ether polyethylene glycol). [47]

Later, in 2021, Pei et al.'s review of carbon capture via electrochemistry provided a collection of data on post-combustion carbon capture. Applied in energy and chemical plants, as well as dispersed point sources, post-combustion capture is currently under development with its widespread application still being explored. Furthermore, electrochemical CO_2 capture is differentiated into redox carrier and pH swing, where carbon is captured and released through transport mechanisms. The redox carrier method uses a redox-active molecule with 2 stable oxidation states to capture CO_2. Applying a reductive current generates a reduced form of the molecule, denoted R^{n-}, with a high binding affinity, allowing it to adsorb CO_2 from a dilute stream and form carbonate adducts. In contrast, applying an oxidative current generates the oxidized form, which releases CO_2 gas, which enables CO_2 capture and concentration. The pH swing method incorporates pH-dependent conversion of carbonaceous species, specifically bicarbonate and carbonate. Acidifying the formed carbonate/bicarbonate solutions using aqueous inorganic alkaline solutions, CO_2 is released. This local pH is manipulated through various electrochemical techniques. Recent advancements have explored the application of electrochemical pH-swing methods to capture carbon from seawater, which can potentially recover calcium carbonate.[48-50] Ultimately, this innovation is facing multiple challenges due to the constant demand for high-purity gaseous CO_2 and the economic viability of the multi-step processes. Current research should aim to overcome these challenges and improve the technology's feasibility for widespread application.[51]

Other decarbonization technologies could also be categorized into demand-side and supply-side approaches. Demand-side decarbonization allows energy efficiency to be increased while reducing demand. Supply-side decarbonization is differentiated into four categories: renewable energies, low-carbon fossil energies, nuclear, and hydrogen. From research, Asian countries showed potential for installation of CCS in coal-fired power plants to mitigate the emitted CO_2.[52] Furthermore, CCS can be utilized in natural gas-fired power plants (GP-CCS), industrial facilities (Ind-CCS), and coal-fired power plants (CP-CCS). The methods of decarbonization explored suggest CCS' ability to become a primary decarbonization pathway across multiple industrial sectors, considering that the method has gained more research attention over the past few years. The paper lastly advises that the target decarbonization rate to be achieved by 2050, estimated for 80 economies, would increase rapidly as we approach 2050. Thus, we should start the implementation of carbon capture technologies, as soon as possible, as delays will only bring more opportunity costs.[53]

Research in 2023 developed a new carbon capture-based organo-mineral fertilizer (CCOMF). This will mineralize and provide nutrients sufficient to satisfy the needs of cereal crops while also demonstrating the potential of utilizing point-source carbon capture technology to create sustainable fertilizers. Overall, the CCOMF produced comparable yields to

conventional mineral fertilizers with no additional short-term negative impacts. Organo-mineral fertilizers (OMF) combine organic waste products with mineral fertilizers to produce a more desirable and balanced nutrient content.[54] Often, CCOMF is produced in different formulations with different percentages of nitrogen and pellets made in batches of varying sizes. This process incorporates point-source carbon dioxide into organic waste materials to capture carbon. Then, CO_2 is mixed into organic waste stream to form ammonium nitrate and calcium carbonate after addition of calcium nitrate and aqueous ammonia.[55] Studies evaluate the efficacy of CCOMF through field trials, which produce comparable crop yields without negative impacts on soil health, highlighting the feasibility of employing carbon capture technology in the production of sustainable fertilizers. Previous observation and research, alongside Burak et al.'s research, suggested that the integration of carbon capture technology into the production of OMFs does not negatively impact crop yields.[56-59]

Moreover, research specifically aimed at post-combustion carbon capture studied cryogenic carbon capture via desublimation in detail. In this process, flue gas is dried to remove moisture, then frozen below freezing temperature of CO_2, allowing the gas to desublimate. The solid CO_2 is subsequently isolated and converted into a liquid state to facilitate carbon utilization. This process has a high overall CO_2 recovery, with some papers reporting up to 99% efficiencies.[60] Additionally, this process requires less energy than traditional carbon capture technologies, with an estimated energy expenditure of 0.63 MJ e/kg CO_2. Compared to amine-based systems, which have an efficiency of 20%, the process has greater efficiencies of over 50%, making it favorable. Along with reducing water and steam usage, the potential for connection with large-scale energy storage systems, and the ability to separate ancillary pollutants, the carbon capture rate is easily above industry standard. However, issues including frost obstructions, material durability at cryogenic temperatures, and expensive equipment continue to plague the cryogenic post-combustion industry, as shown in Figure 4. Recent developments include dynamic packed ped technology and external cooling loops to enhance efficiency and address frost-clogging issues.[61,62]

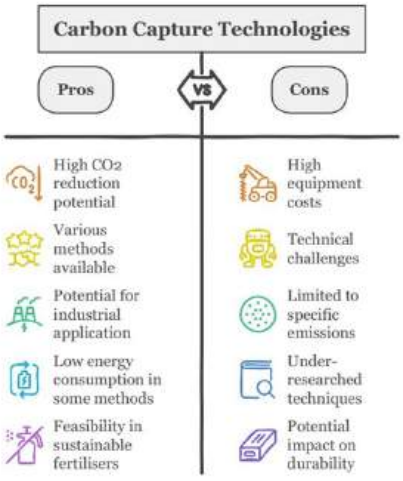


Figure 4: A comparison of advantages and disadvantages associated with carbon capture technologies

2.2 Direct Air Capture

Direct Air Capture (DAC) is a carbon capture technique that removes carbon from the surrounding air. The process involves sorbents, either solid or liquid, to extract carbon using contractors with large surface area to isolate, release and store carbon, which has been proven to aid with the capture of carbon from scattered emission sources. To have a meaningful impact by 2050, DAC needs to become a cost-effective option that has expanded at a global scale.[63] Figure 5 presents innovative advancements in DAC technologies in recent years.

A review of climate mitigation strategies rightfully summarizes the negative emission technologies, including BECCS, biochar, enhanced weathering, DAC, ocean fertilization, ocean alkalinity enhancement, soil carbon sequestration, and more. The fundamental concept of Direct air carbon capture and storage (DACCS) is the removal of atmospheric CO_2 through chemical bonding in sorbents, which is then stored in geological reservoirs or used for industrial reasons. The two processes are adsorption, where CO_2 forms a bond with the solid sorbents, and absorption, where CO_2 dissolves in the liquid sorbents. In addition to the energy needed to power fans, pumps, and compressors, the procedure also needs heat energy for sorbent regeneration. Publications emphasize the significant disadvantage of high related expenses and the risk to the integrity of CO_2 storage.[64,65] With an anticipated carbon removal range of 0.5-5 $GtCO_2year^{-1}$, DAC requires three times as much energy per ton as conventional carbon capture.[66,67] Another advantage is that it can be located anywhere.[68]

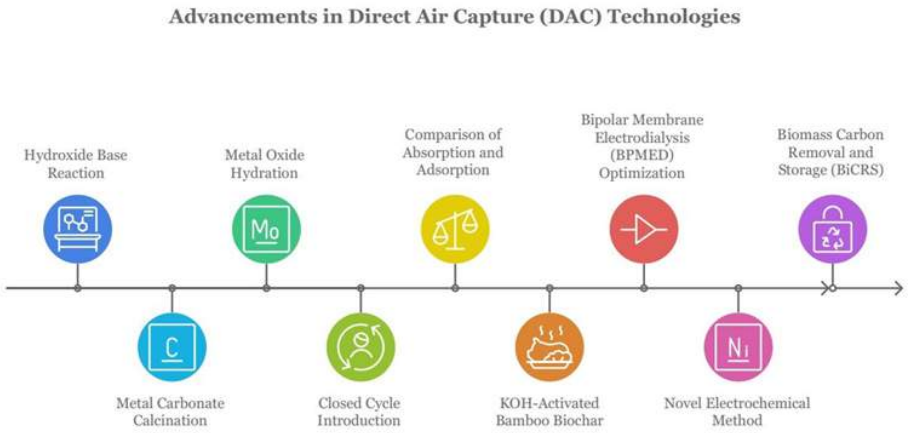


Figure 5: Advancements of Direct Air Capture technologies reviewed in this literature

In contrast to point source capture, which is useful in industry to produce power and heat, DAC can absorb carbon emissions across numerous sectors, according to a 2021 assessment of porous material by Siegelman et al. The research discusses that amine-functionalized materials, such as porous polymers, metal-organic frameworks, and silicas, provide the required CO_2 selectivity at lower partial pressures to enable DAC despite reviewing various porous materials. Organic monomers are used to create porous organic networks, which have a high degree of chemical tunability.[69] Although CO_2-targeting functional groups, such as amines, can be directly

integrated into the polymer backbone through covalent connections, these frameworks still need to be developed and studied further before they can be used on an industrial scale. Inorganic ions joined by organic ligands produce metal-organic frameworks. Because of their diverse chemical composition and shape, their porous structure can be chemically modified for use in gas separation, which is utilized in CCS.[70] Lastly, amine-functionalized silica uses selected pore volumes to offer more active sites for the amine-CO_2 reaction, facilitating adsorption.[71,72]

Carbon capture technologies, specifically DAC, in air conditioning, are currently under research, and the hope is to integrate the two technologies to improve air quality, save energy, and address climate change. Present-day air conditioning (AC) systems lower the temperature of the air they filter from their surroundings and then release it into the structure. Current filtration techniques only ensure that dust is eliminated from the hot air, leaving the original constituent parts unaltered. Whilst some have attempted to actively incorporate DAC into the architecture of building HVAC systems, others have tested the adsorption capabilities of adsorbents indoors.[73,74] An interesting ventilation arrangement proposed by Kim and Leibundgut recommended the use of a low air recirculation ratio coupled with CO_2 capture devices to reduce energy consumption in HVAC systems.[75] Further research has shown that the integration of DAC with HVAC systems leads to 30-60% energy savings.[76] Sorbents that hold major potential for implementation are nanoscale carbon nitride and metal-organic framework: they are both environmentally friendly, with high adsorptive uptake of CO_2. Successful outcomes in these research efforts have drawn the conclusion that AC units with CO_2 capture could bring a major change if widely implemented.[77]

Technologies utilizing DAC are categorized into adsorption, absorption, ion exchange resin, mineral carbonation, photocatalysis, cryogenic separation, electrochemical approaches, and electrodialysis approaches. Specifically, absorption and adsorption are heavily researched in the industry. Aqueous solutions of strong bases that include hydroxides such as NaOH or KOH are utilized for absorption. The absorption process proposed by Lackner et al. follows a set of 3 reactions.[78]

$$Ca(OH)_2 + CO_2 \rightarrow CaCO_3 + H_2O \quad (1)$$

$$CaCO_3 \rightarrow CaO + CO_2 \quad (2)$$

$$CaO + H_2O \rightarrow Ca(OH)_2 \quad (3)$$

From these recurring reactions, the hydroxide base reacts with CO_2 to form a metal carbonate and water. This metal carbonate undergoes calcination to form metal oxide and CO_2, where the metal oxide is later hydrated to produce the original hydroxide base. Introducing a closed cycle, absorption is applicable to other hydroxide bases with some alternate arrangements. In contrast, adsorption relies on sorbents: chemisorbents are largely proposed throughout research, whereas physisorbents are still under investigation due to their unique structures and behavior.

Adsorption and absorption have their respective advantages and disadvantages. Absorption requires base solvents to be present up to 30% wt, limiting the amount of interaction

with CO_2. This process has highly complex regeneration despite working continuously and requiring high temperatures for regeneration supplied by high-grade heat sources. Adsorption technology offers significant modularity with compact capture facilities, utilizing cyclic processes that depend on changes in temperature or pressure at the installation site, which could impact cost due to the retrofit of the system. However, it is almost impossible to quantify which process is more suitable in general since context must be applied to reach a conclusion [79]

Employing a new sorbent material, Zhang et al.'s research used KOH-activated bamboo biochar as a sorbent for DAC. The activation of bamboo charcoal was physically mixed with potassium hydroxide KOH in ratios 1:1, 1:0.5, and 1:0.2. This is carried out to improve the morphology and texture of biochar, which increases its surface area. This was powdered and heated up to 700°C, where compounds containing potassium were washed out with deionized water and hydrochloride until the filtrate was neutral. These were then dried overnight and used as biochar samples. As the activation degree increases, the equilibrium of the adsorption process is slowed down. This is determined by bamboo biochar mixed with KOH in a 1:1 ratio, which has the highest adsorption time and exhibits the highest capacity of CO_2 uptake. The paper concludes that bamboo biochar serves as a promising solid adsorbent for DAC. One caution that must be taken is that the biochar shows stable sorbent ability under dry conditions, whereas higher humidity can reduce adsorption abilities by 36%.[80]

Innovative analysis of Bipolar Membrane Electrodialysis (BPMED) technology discussed the optimization of BPMED systems with existing DAC methods. In this process, CO_2 is captured through chemical absorption in KOH, which reacts to form K_2CO_3. The BPMED unit is used for regeneration, where its efficiency is enhanced through modeling and optimization. There is a higher regeneration ratio and efficiency than other methods. This leaves opportunities for cost reduction, as the decreasing cost of electricity from renewable sources can also lower the operating costs for BPMED processes. However, there are high energy demands and a high cost of membranes, with a short lifespan of three years, meaning that this requires frequent, costly replacements. The paper recommends that future advancements focus on enhancing electrical conductivity and overall stability, as power consumption and membrane replacement significantly impact the economics of the process.[81]

An electrochemical method for DAC utilizing neutral red as a redox-active substance enhanced by nicotinamide was presented in Seo and Hatton's research. This method significantly reduces energy requirements for carbon capture: neutral red is used in its reduced form, known as NRH2. The solution is saturated and acidified once ambient air is bubbled through, making the solution become more neutral from pH 12 to pH 9.1. The CO_2-saturated solution is electrochemically oxidized, which regenerates neutral red (NR) from NRH2, releasing free CO_2 back into the atmosphere. Operating in continuous flow, its effectiveness is assessed by quantifying the amount of CO_2 emitted and the utilization of electrons throughout the process. The energy requirements are estimated to be 35kJ e/mol for a 15% CO_2 feed and 65kJ e/mol for direct air capture.[82]

Biomass carbon removal and storage (BiCRS) refers to a collection of carbon dioxide removal technologies that depend on photosynthesis, followed by the stabilization and sequestration of biomass carbon. The leading methods of BiCRS include gasification, pyrolysis, combustion, anaerobic digestion, fermentation, and biomass burial. It is classified as DAC due to its mechanism of extracting carbon from ambient air. Woods et al. introduce composting with CO_2 capture. This process begins with CO_2 fixation in living biomass through photosynthesis, followed by temporary carbon storage such as forestry or agricultural residues, as well as animal or municipal waste. Composting allows for the passive conversion of the carbon stored to CO_2 via aerobic respiration. This is an autothermal process in which the operating temperature is kept between 40-65°C. This method identifies pre- and post-combusting CO_2 capture, where the gas separation takes place either prior to or after composting via CO_2 separation from N_2. The results demonstrate a high capture of high-purity gaseous CO_2, with concentrations ranging from 18-95 volume percent, with up to 98% less energy for final CO_2 capture and purification.[83]

Conclusion

In conclusion, the industrial implementation of CCS remains unclear. Evaluation of conventional and innovative carbon capture technologies has shown that DAC is seen as the most viable option to mitigate climate change. Point Source Carbon capture is effective due to its potential to extract large volumes of carbon but solely relies on costly installation on operational generation sources. Direct Air Capture, on the other hand, can extract carbon out of ambient air but is associated with high costs and CO_2 storage integrity. Multiple research projects have shown that DAC has a higher potential to mitigate climate change in hopes of meeting the Paris Agreement's goal of limiting global warming to 1.5°C or below 2°C.[84] Despite current trends focused on carbon utilization, we suggest that there should also be a focus on the large-scale deployment of carbon capture technologies globally to rapidly work towards the Paris Agreement's objectives by 2050.

Works Cited

1. Topham, Susan, et al. Carbon Dioxide. 2014, pp. 1–43. ResearchGate, https://doi.org/10.1002/14356007.a05_165.pub2.
2. Yoro, Kelvin O., and Michael O. Daramola. "CO_2 Emission Sources, Greenhouse Gases, and the Global Warming Effect." Advances in Carbon Capture, edited by Mohammad Reza Rahimpour et al., Woodhead Publishing, 2020, pp. 3–28. ScienceDirect, https://doi.org/10.1016/B978-0-12-819657-1.00001-3.
3. Environmental Impact Assessment - European Commission. https://environment.ec.europa.eu/law-and-governance/environmental-assessments/environmental-impact-assessment_en. Accessed 7 Aug. 2024.
4. Giles, A. Operations Manual: Threshold Limit Values (TLV) For Chemical Substances Committee, 2020. American Conference of Governmental Industrial Hygienists. https://www.acgih.org/wp-content/uploads/2021/02/Operations-Manual-TLV-CS_11-7-2020.pdf (accessed 2024-08-28).
5. Franco, Alessandro, and Francesco Leccese. "Measurement of CO_2 Concentration for Occupancy Estimation in Educational Buildings with Energy Efficiency Purposes." Journal of Building Engineering, vol. 32, Nov. 2020, p. 101714. PubMed Central, https://doi.org/10.1016/j.jobe.2020.101714.
6. The Paris Agreement | UNFCCC. https://unfccc.int/process-and-meetings/the-paris-agreement. Accessed 7 Aug. 2024.
7. Ekemezie, Ifeanyi Onyedika, and Wags Numoipiri Digitemie. "CARBON CAPTURE AND UTILIZATION (CCU): A REVIEW OF EMERGING APPLICATIONS AND CHALLENGES." Engineering Science & Technology Journal, vol. 5, no. 3, 3, Mar. 2024, pp. 949–61. www.fepbl.com, https://doi.org/10.51594/estj.v5i3.949.
8. Ma, Jinfeng, et al. "Carbon Capture and Storage: History and the Road Ahead." Engineering, vol. 14, July 2022, pp. 33–43. ScienceDirect, https://doi.org/10.1016/j.eng.2021.11.024.
9. Sodiq, Ahmed, et al. "A Review on Progress Made in Direct Air Capture of CO_2" Environmental Technology & Innovation, vol. 29, Feb. 2023, p. 102991. ScienceDirect, https://doi.org/10.1016/j.eti.2022.102991.
10. "CCS Image Library." Global CCS Institute, https://www.globalccsinstitute.com/resources/ccs-image-library/. Accessed 8 Aug. 2024.
11. Abanades, J. Carlos, et al. "An Air CO_2 Capture System Based on the Passive Carbonation of Large Ca(OH)2 Structures." Sustainable Energy & Fuels, vol. 4, no. 7, June 2020, pp. 3409–17. pubs.rsc.org, https://doi.org/10.1039/D0SE00094A.
12. Ren, Lei, et al. "A Review of CO_2 Emissions Reduction Technologies and Low-Carbon Development in the Iron and Steel Industry Focusing on China." Renewable and Sustainable Energy Reviews, vol. 143, June 2021, p. 110846. ScienceDirect, https://doi.org/10.1016/j.rser.2021.110846.

13. Boot-Handford, Matthew E., et al. “Carbon Capture and Storage Update.” Energy & Environmental Science, vol. 7, no. 1, Dec. 2013, pp. 130–89. pubs.rsc.org, https://doi.org/10.1039/C3EE42350F.
14. Al-Mamoori, Ahmed, et al. “Carbon Capture and Utilization Update.” Energy Technology, vol. 5, no. 6, 2017, pp. 834–49. Wiley Online Library, https://doi.org/10.1002/ente.201600747.
15. Ekemezie, Ifeanyi Onyedika, and Wags Numoipiri Digitemie. “CARBON CAPTURE AND UTILIZATION (CCU): A REVIEW OF EMERGING APPLICATIONS AND CHALLENGES.” Engineering Science & Technology Journal, vol. 5, no. 3, 3, Mar. 2024, pp. 949–61. www.fepbl.com, https://doi.org/10.51594/estj.v5i3.949.
16. Williams, James H., et al. “Carbon-Neutral Pathways for the United States.” AGU Advances, vol. 2, no. 1, 2021, p. e2020AV000284. Wiley Online Library, https://doi.org/10.1029/2020AV000284.
17. “DOE Industrial Decarbonization Roadmap.” Energy.Gov, https://www.energy.gov/industrial-technologies/doe-industrial-decarbonization-roadmap. Accessed 25 Aug. 2024.
18. “How Energy Efficiency Will Power Net Zero Climate Goals – Analysis.” IEA, 29 Mar. 2021, https://www.iea.org/commentaries/how-energy-efficiency-will-power-net-zero-climate-goals.
19. Bui, Mai, et al. “Carbon Capture and Storage (CCS): The Way Forward.” Energy & Environmental Science, vol. 11, no. 5, 2018, pp. 1062–176. pubs.rsc.org, https://doi.org/10.1039/C7EE02342A.
20. Freer, Muir, et al. “Putting Bioenergy With Carbon Capture and Storage in a Spatial Context: What Should Go Where?” Frontiers in Climate, vol. 4, Mar. 2022. Frontiers, https://doi.org/10.3389/fclim.2022.826982.
21. Allohverdi, Tara, et al. “A Review on Current Status of Biochar Uses in Agriculture.” Molecules, vol. 26, no. 18, 18, Jan. 2021, p. 5584. www.mdpi.com, https://doi.org/10.3390/molecules26185584.
22. Yaashikaa, P. R., et al. “A Critical Review on the Biochar Production Techniques, Characterization, Stability and Applications for Circular Bioeconomy.” Biotechnology Reports, vol. 28, Dec. 2020, p. e00570. ScienceDirect, https://doi.org/10.1016/j.btre.2020.e00570.
23. Zhang, Chen, et al. “The Application of Biochar for CO_2 Capture: Influence of Biochar Preparation and CO_2 Capture Reactors.” Industrial & Engineering Chemistry Research, vol. 62, no. 42, Oct. 2023, pp. 17168–81. ACS Publications, https://doi.org/10.1021/acs.iecr.3c00445.
24.Global Status of CCS. https://status22.globalccsinstitute.com/2022-status-report/global-status-of-ccs/. Accessed 26 Aug. 2024.

25. Madejski, Paweł, et al. “Methods and Techniques for CO_2 Capture: Review of Potential Solutions and Applications in Modern Energy Technologies.” Energies, vol. 15, Jan. 2022, p. 887. ResearchGate, https://doi.org/10.3390/en15030887.
26. Voldsund, Mari, et al. “Low Carbon Power Generation for Offshore Oil and Gas Production.” Energy Conversion and Management: X, vol. 17, Jan. 2023, p. 100347. ScienceDirect, https://doi.org/10.1016/j.ecmx.2023.100347.
27. Madejski, Paweł, et al. “Methods and Techniques for CO_2 Capture: Review of Potential Solutions and Applications in Modern Energy Technologies.” Energies, vol. 15, Jan. 2022, p. 887. ResearchGate, https://doi.org/10.3390/en15030887.
28. Dubey, Aseem, and Akhilesh Arora. “Advancements in Carbon Capture Technologies: A Review.” Journal of Cleaner Production, vol. 373, Nov. 2022, p. 133932. ScienceDirect, https://doi.org/10.1016/j.jclepro.2022.133932.
29. Ibigbami, Olayinka Abidemi, et al. “Post-Combustion Capture and Other Carbon Capture and Sequestration (CCS) Technologies: A Review.” Environmental Quality Management, vol. 34, no. 1, 2024, p. e22180. Wiley Online Library, https://doi.org/10.1002/tqem.22180.
30. Madejski, Paweł, et al. “Methods and Techniques for CO_2 Capture: Review of Potential Solutions and Applications in Modern Energy Technologies.” Energies, vol. 15, Jan. 2022, p. 887. ResearchGate, https://doi.org/10.3390/en15030887.
31. Talei, Saeed, et al. “Oxyfuel Combustion Makes Carbon Capture More Efficient.” ACS Omega, vol. 9, no. 3, Jan. 2024, pp. 3250–61. ACS Publications, https://doi.org/10.1021/acsomega.3c05034.
32. Bazooyar, Bahamin, and Abolfazl Jomekian. “Chapter Twenty-One - Oxyfuel Combustion as a Carbon Capture Technique.” Advances and Technology Development in Greenhouse Gases: Emission, Capture and Conversion, edited by Mohammad Reza Rahimpour et al., Elsevier, 2024, pp. 437–95. ScienceDirect, https://doi.org/10.1016/B978-0-443-19233-3.00023-7.
33. Maroto-Valer, M. Mercedes, et al. “CO_2 Capture by Activated and Impregnated Anthracites.” Fuel Processing Technology, vol. 86, no. 14, Oct. 2005, pp. 1487–502. ScienceDirect, https://doi.org/10.1016/j.fuproc.2005.01.003.
34. Stewart, Caleb, and Mir-Akbar Hessami. “A Study of Methods of Carbon Dioxide Capture and Sequestration—the Sustainability of a Photosynthetic Bioreactor Approach.” Energy Conversion and Management, vol. 46, no. 3, Feb. 2005, pp. 403–20. ScienceDirect, https://doi.org/10.1016/j.enconman.2004.03.009.
35. Powell, Clem E., and Greg G. Qiao. “Polymeric CO_2/N_2 Gas Separation Membranes for the Capture of Carbon Dioxide from Power Plant Flue Gases.” Journal of Membrane Science, vol. 279, no. 1, Aug. 2006, pp. 1–49. ScienceDirect, https://doi.org/10.1016/j.memsci.2005.12.062.
36. Rocha, Luis A. M., et al. “Separation of CO_2/CH_4 Using Carbon Molecular Sieve (CMS) at Low and High Pressure.” Chemical Engineering Science, vol. 164, June 2017, pp. 148–57. ScienceDirect, https://doi.org/10.1016/j.ces.2017.01.071.

37. Brinker, C. J., et al. “Sol-Gel Strategies for Controlled Porosity Inorganic Materials.” Journal of Membrane Science, vol. 94, no. 1, Sept. 1994, pp. 85–102. ScienceDirect, https://doi.org/10.1016/0376-7388(93)E0129-8.
38. Mikulčić, Hrvoje, et al. “Flexible Carbon Capture and Utilization Technologies in Future Energy Systems and the Utilization Pathways of Captured CO_2.” Renewable and Sustainable Energy Reviews, vol. 114, Oct. 2019, p. 109338. ScienceDirect, https://doi.org/10.1016/j.rser.2019.109338.
39. Garcia, Jose Antonio, et al. “Technical Analysis of CO_2 Capture Pathways and Technologies.” Journal of Environmental Chemical Engineering, vol. 10, no. 5, Oct. 2022, p. 108470. ScienceDirect, https://doi.org/10.1016/j.jece.2022.108470.
40. Habert, G., et al. “Environmental Impacts and Decarbonization Strategies in the Cement and Concrete Industries.” Nature Reviews Earth & Environment, vol. 1, no. 11, Sept. 2020, pp. 559–73. Semantic Scholar, https://doi.org/10.1038/s43017-020-0093-3.
41. Raj, John, and A. Seetharaman. “Role of Waste and Performance Management in the Construction Industry.” Journal of Environmental Science and Technology, vol. 6, Mar. 2013, pp. 119–29. ResearchGate, https://doi.org/10.3923/jest.2013.119.129.
42. Ho, Minh T., et al. “Comparison of CO_2 Capture Economics for Iron and Steel Mills.” International Journal of Greenhouse Gas Control, vol. 19, Nov. 2013, pp. 145–59. ScienceDirect, https://doi.org/10.1016/j.ijggc.2013.08.003.
43. Ren, Lei, et al. “A Review of CO_2 Emissions Reduction Technologies and Low-Carbon Development in the Iron and Steel Industry Focusing on China.” Renewable and Sustainable Energy Reviews, vol. 143, June 2021, p. 110846. ScienceDirect, https://doi.org/10.1016/j.rser.2021.110846
44. Raganati, Federica, et al. “Adsorption of Carbon Dioxide for Post-Combustion Capture: A Review.” Energy & Fuels, vol. 35, no. 16, Aug. 2021, pp. 12845–68. ACS Publications, https://doi.org/10.1021/acs.energyfuels.1c01618.
45. Mondal, Monoj Kumar, et al. “Progress and Trends in CO_2 Capture/Separation Technologies: A Review.” Energy, vol. 46, no. 1, Oct. 2012, pp. 431–41. ScienceDirect, https://doi.org/10.1016/j.energy.2012.08.006.
46. Pellegrini, Laura A., et al. “Energy Saving in a CO_2 Capture Plant by MEA Scrubbing.” Chemical Engineering Research and Design, vol. 89, no. 9, Sept. 2011, pp. 1676–83. ScienceDirect, https://doi.org/10.1016/j.cherd.2010.09.024.
47. Garcia, Jose Antonio, et al. “Technical Analysis of CO_2 Capture Pathways and Technologies.” Journal of Environmental Chemical Engineering, vol. 10, no. 5, Oct. 2022, p. 108470. ScienceDirect, https://doi.org/10.1016/j.jece.2022.108470.
48. Mowbray, Benjamin A. W., et al. “Electrochemical Cement Clinker Precursor Production at Low Voltages.” ACS Energy Letters, vol. 8, no. 4, Apr. 2023, pp. 1772–78. ACS Publications, https://doi.org/10.1021/acsenergylett.3c00242.

49. Yan, Litao, et al. “An Electrochemical Hydrogen-Looping System for Low-Cost CO_2 Capture from Seawater.” ACS Energy Letters, vol. 7, no. 6, June 2022, pp. 1947–52. ACS Publications, https://doi.org/10.1021/acsenergylett.2c00396.
50. Kim, Seoni, et al. “Asymmetric Chloride-Mediated Electrochemical Process for CO_2 Removal from Oceanwater.” Energy & Environmental Science, vol. 16, no. 5, May 2023, pp. 2030–44. pubs.rsc.org, https://doi.org/10.1039/D2EE03804H.
51. Pei, Yuhou, et al. “Carbon Capture and Utilization via Electrochemistry, What’s next?” Next Nanotechnology, vol. 3–4, Sept. 2023, p. 100020. ScienceDirect, https://doi.org/10.1016/j.nxnano.2023.100020.
52. Lau, Hon Chung. “The Contribution of Carbon Capture and Storage to the Decarbonization of Coal-Fired Power Plants in Selected Asian Countries.” Energy & Fuels, vol. 37, no. 20, Oct. 2023, pp. 15919–34. ACS Publications, https://doi.org/10.1021/acs.energyfuels.3c02648.
53. Lau, Hon Chung, and Steve C. Tsai. “Global Decarbonization: Current Status and What It Will Take to Achieve Net Zero by 2050.” Energies, vol. 16, no. 23, Nov. 2023, p. 7800. Semantic Scholar, https://doi.org/10.3390/en16237800.
54. Lake, J. A., et al. “Sustainable Soil Improvement and Water Use in Agriculture: CCU Enabling Technologies Afford an Innovative Approach.” Journal of CO_2 Utilization, vol. 32, July 2019, pp. 21–30. ScienceDirect, https://doi.org/10.1016/j.jcou.2019.03.010.
55. Burak, Emma, and Ruben Sakrabani. “Novel Carbon Capture-Based Organo-Mineral Fertilisers Show Comparable Yields and Impacts on Soil Health to Mineral Fertiliser across Two Cereal Crop Field Trials in Eastern England.” Field Crops Research, vol. 302, Oct. 2023, p. 109043. ScienceDirect, https://doi.org/10.1016/j.fcr.2023.109043.
56. Pawlett, M., et al. “Nutrient Potential of Biosolids and Urea Derived Organo-Mineral Fertilisers in a Field Scale Experiment Using Ryegrass (Lolium Perenne L.).” Field Crops Research, vol. 175, Apr. 2015, pp. 56–63. ScienceDirect, https://doi.org/10.1016/j.fcr.2015.02.006.
57. Deeks, Lynda K., et al. “A New Sludge-Derived Organo-Mineral Fertilizer Gives Similar Crop Yields as Conventional Fertilizers.” Agronomy for Sustainable Development, vol. 33, no. 3, July 2013, pp. 539–49. Springer Link, https://doi.org/10.1007/s13593-013-0135-z.
58. Antille, Diogenes L., et al. “Field-Scale Evaluation of Biosolids-Derived Organomineral Fertilizers Applied to Winter Wheat in England.” Agronomy Journal, vol. 109, no. 2, 2017, pp. 654–74. Wiley Online Library, https://doi.org/10.2134/agronj2016.09.0495.
59. Burak, Emma, and Ruben Sakrabani. “Novel Carbon Capture-Based Organo-Mineral Fertilisers Show Comparable Yields and Impacts on Soil Health to Mineral Fertiliser across Two Cereal Crop Field Trials in Eastern England.” Field Crops Research, vol. 302, Oct. 2023, p. 109043. ScienceDirect, https://doi.org/10.1016/j.fcr.2023.109043.
60. Baxter, Larry, et al. Cryogenic Carbon Capture Development Final/Technical Report. 2019, p. DOE-SES-28697, 1572908. Semantic Scholar, https://doi.org/10.2172/1572908.

61. Tuinier, M. J., et al. “Techno-Economic Evaluation of Cryogenic CO_2 Capture—A Comparison with Absorption and Membrane Technology.” International Journal of Greenhouse Gas Control, vol. 5, no. 6, Nov. 2011, pp. 1559–65. ScienceDirect, https://doi.org/10.1016/j.ijggc.2011.08.013.
62. Luberti, Mauro, et al. “Unveiling the Potential of Cryogenic Post-Combustion Carbon Capture: From Fundamentals to Innovative Processes.” Energies, vol. 17, no. 11, June 2024, p. 2673. research.manchester.ac.uk, https://doi.org/10.3390/en17112673.
63. Ozkan, Mihrimah, et al. “Current Status and Pillars of Direct Air Capture Technologies.” iScience, vol. 25, no. 4, Apr. 2022, p. 103990. ScienceDirect, https://doi.org/10.1016/j.isci.2022.103990.
64. Fuss, Sabine, et al. “Negative Emissions—Part 2: Costs, Potentials and Side Effects.” Environmental Research Letters, vol. 13, no. 6, May 2018, p. 063002. Institute of Physics, https://doi.org/10.1088/1748-9326/aabf9f.
65. Azapagic, Adisa, et al. “Report: Greenhouse Gas Removal,” Royal Society, https://royalsociety.org/-/media/policy/projects/greenhouse-gas-removal/royal-society-greenhouse-gas-removal-report-2018.pdf Accessed 29 Sep. 2024.
66. Gambhir, Ajay, and Massimo Tavoni. “Direct Air Carbon Capture and Sequestration: How It Works and How It Could Contribute to Climate-Change Mitigation.” One Earth, vol. 1, no. 4, Dec. 2019, pp. 405–09. www.cell.com, https://doi.org/10.1016/j.oneear.2019.11.006.
67. Fuss, Sabine, et al. “Negative Emissions—Part 2: Costs, Potentials and Side Effects.” Environmental Research Letters, vol. 13, no. 6, May 2018, p. 063002. Institute of Physics, https://doi.org/10.1088/1748-9326/aabf9f.
68. Fawzy, Samer, et al. “Strategies for Mitigation of Climate Change: A Review.” Environmental Chemistry Letters, vol. 18, July 2020. ResearchGate, https://doi.org/10.1007/s10311-020-01059-w.
69. Zeng, Yongfei, et al. “Covalent Organic Frameworks for CO_2 Capture.” Advanced Materials, vol. 28, no. 15, 2016, pp. 2855–73. Wiley Online Library, https://doi.org/10.1002/adma.201505004.
70. Furukawa, Hiroyasu, et al. “The Chemistry and Applications of Metal-Organic Frameworks.” Science (New York, N.Y.), vol. 341, no. 6149, Aug. 2013, p. 1230444. PubMed, https://doi.org/10.1126/science.1230444.
71. Girimonte, Rossella, et al. “Amine-Functionalized Mesoporous Silica Adsorbent for CO_2 Capture in Confined-Fluidized Bed: Study of the Breakthrough Adsorption Curves as a Function of Several Operating Variables.” Processes, vol. 10, no. 2, 2, Feb. 2022, p. 422. www.mdpi.com, https://doi.org/10.3390/pr10020422.
72. Siegelman, Rebecca L., et al. “Porous Materials for Carbon Dioxide Separations.” Nature Materials, vol. 20, no. 8, Aug. 2021, pp. 1060–72. www.nature.com, https://doi.org/10.1038/s41563-021-01054-8.
73. Zhao, Ruikai, et al. “Thermodynamic Exploration of Temperature Vacuum Swing Adsorption for Direct Air Capture of Carbon Dioxide in Buildings.” Energy Conversion and

Management, vol. 183, Mar. 2019, pp. 418–26. ScienceDirect, https://doi.org/10.1016/j.enconman.2019.01.009.
74. Gall, Elliott T., et al. “Investigating CO_2 Removal by Ca- and Mg-Based Sorbents with Application to Indoor Air Treatment.” Building and Environment, vol. 110, Dec. 2016, pp. 161–72. ScienceDirect, https://doi.org/10.1016/j.buildenv.2016.10.008.
75. Kim, Moon Keun, and Hansjürg Leibundgut. “Performance of Novel Ventilation Strategy for Capturing CO_2 with Scheduled Occupancy Diversity and Infiltration Rate.” Building and Environment, vol. 89, July 2015, pp. 318–26. ScienceDirect, https://doi.org/10.1016/j.buildenv.2015.02.012.
76. Kim, Moon Keun, et al. “A Novel Ventilation Strategy with CO_2 Capture Device and Energy Saving in Buildings.” Energy and Buildings, vol. 87, Jan. 2015, pp. 134–41. ScienceDirect, https://doi.org/10.1016/j.enbuild.2014.11.017.
77. Sodiq, Ahmed, et al. “A Review on Progress Made in Direct Air Capture of CO_2.” Environmental Technology & Innovation, vol. 29, Feb. 2023, p. 102991. ScienceDirect, https://doi.org/10.1016/j.eti.2022.102991.
78. Lackner, K. S. “Capture of Carbon Dioxide from Ambient Air.” The European Physical Journal Special Topics, vol. 176, no. 1, Sept. 2009, pp. 93–106. Springer Link, https://doi.org/10.1140/epjst/e2009-01150-3.
79. Leonzio, Grazia, et al. “Analysis of Technologies for Carbon Dioxide Capture from the Air.” Applied Sciences, vol. 12, no. 16, 16, Jan. 2022, p. 8321. www.mdpi.com, https://doi.org/10.3390/app12168321.
80. Zhang, Chen, et al. “Direct Air Capture of CO_2 by KOH-Activated Bamboo Biochar.” Journal of the Energy Institute, vol. 105, Dec. 2022, pp. 399–405. ScienceDirect, https://doi.org/10.1016/j.joei.2022.10.017.
81. Sabatino, Francesco, et al. “Modeling, Optimization, and Techno-Economic Analysis of Bipolar Membrane Electrodialysis for Direct Air Capture Processes.” Industrial & Engineering Chemistry Research, vol. 61, no. 34, Aug. 2022, pp. 12668–79. ACS Publications, https://doi.org/10.1021/acs.iecr.2c00889.
82. Seo, Hyowon, and T. Alan Hatton. “Electrochemical Direct Air Capture of CO_2 Using Neutral Red as Reversible Redox-Active Material.” Nature Communications, vol. 14, no. 1, Jan. 2023, p. 313. www.nature.com, https://doi.org/10.1038/s41467-023-35866-w.
83. Woods, Ethan, et al. “Biomass Composting with Gaseous Carbon Dioxide Capture.” RSC Sustainability, vol. 2, no. 3, Mar. 2024, pp. 621–25. pubs.rsc.org, https://doi.org/10.1039/D3SU00411B.
84. Huang, Meng-Tian, and Pan-Mao Zhai. “Achieving Paris Agreement Temperature Goals Requires Carbon Neutrality by Middle Century with Far-Reaching Transitions in the Whole Society.” Advances in Climate Change Research, vol. 12, no. 2, Apr. 2021, pp. 281–86. ScienceDirect, https://doi.org/10.1016/j.accre.2021.03.004.

The History of the American Anti-Abortion Movement in the 19th and 20th Centuries
By Grace Koch

In the current world, abortion is undeniably a highly controversial and debatable topic in America. However, in colonial and early America, abortion was a common, socially acceptable measure that was used by many women(Dine). Common herbs found in household gardens such as savin, tansy, and pennyroyal were used to "clear 'obstructions'"(Blakemore) or "restore regularity"(Abortion) to the menstrual cycle(Peterson). At this time, surgical abortions were uncommon and generally unsafe, but were done in some circumstances(Holland). These attitudes towards abortion were brought from England by the Puritans, who, unlike Catholics, did not see having children as a religious duty and allowed women to choose whether or not they would become mothers(Dine). This was partially due to the widespread belief that the fetus was not alive until the "quickening," or when the pregnant women first felt the baby kick(Holland). Christians believed that this was when "ensoulment" occurred, and any abortions after that point were generally considered morally and religiously wrong("Abortion"). However, the vigorous and influential anti-abortion movement present today would not truly be set in motion until the mid 19th century.

Social and legal positions have changed greatly since this period. This is largely due to the anti-abortion movement and the changes it has gone through. This essay will examine the development of the anti-abortion movement in the United States in the 19th and 20th centuries. First, it will look at the development of social, moral, and religious views on abortion from an accepted option for women to an intolerable act, specifically among those advocating against it. Then, it will evaluate the shifting demographics of the anti-abortion movement from a largely physician-led movement in the Victorian era to a largely religiously motivated movement in the late 20th century. Finally, it will explore the development of medical practices in gynecology and obstetrics and how these affected the anti-abortion movement. These changes continue to shape the modern anti-abortion movement today, as well as the continually shifting abortion laws.

Firstly, social, moral, and religious views of abortion by the general public shifted greatly throughout America's history as its acceptability changed with the popularity of the anti-abortion movement. As mentioned earlier, abortion was extremely common in early America and accepted socially. In the mid 1800s, a woman named Sarah Grosvenor died in a late-term surgical abortion, after the quickening had occurred. She had attempted to abort the fetus using herbs prior to the surgery, but these efforts were unsuccessful. The doctor was then taken to court for killing both Gorsvenor and her baby. Court records show that the women of her community were aware of and accepted her efforts to terminate her pregnancy(Dine). However, there were a select few, mainly religious, who believed abortion to be egregiously immoral. Some went as far as claiming that drugs to alleviate the pain of childbirth were wrong because it "a curse upon women[,] and suffering was necessary to induce maternal love"(Floyd). Aside from the extremists, most men chose to ignore maternal health, including abortions because it did not interfere with their own lives.

In the Victorian era, people were very concerned with modesty, and women were viewed as weak and senseless(Floyd). While these misogynistic ideals were certainly present in previous eras, it peaked in the mid 1800s(Dine). Women were forced into the roles of wife and mother even more vigorously than before(Barrett). As a result there was more pressure on women to stay in the household rather than working(Barret). Women without an independent income needed to rely on a husband or father for money. As a result, these male relatives were more aware of their wife or daughter's reproductive health, or at least the money they were spending for it.

In addition, the Victorian era was a time of sexual repression, which led to women's bodies, as well as their sexual behavior, to be viewed as shameful, especially sex before marriage(Floyd). Abortions were seen as a way to hide this and other improper acts(Lindsey, "*What Did the Suffragists*"). This was partially caused by the general lack of information regarding sexual health, including abortion and birth control, not being available to the general public as a result of censorship laws(Blakemore). The sentiments of shame and lack of education regarding maternal health led to the general public becoming considerably more hostile towards the idea of abortion(Withycombe). Women who got an abortion were vilified in the public eye and were seen as having betrayed their natural, as well as social role, of motherhood(Dine). Despite this, medicinal abortions provided by midwives were still relatively common(Floyd). It is estimated by some historians that around 35% of pregnancies were terminated in the 19th century(Blakemore).

Another factor that contributed to shifting social attitudes towards abortion was the increasing popularity of nativism. During the Victorian era, the birthrate of white Protestants was decreasing due to urbanization("Abortion") and the increased social prevalence of sexual repression(Barrett). Nativists feared that if white, Protestant, American women kept getting abortions, the population of these groups would decline and they would lose power("Abortion"). They wanted to further their agenda by forcing these women to give birth, and convincing them it was morally wrong for them to abort a white fetus.

The most vocal and organized religious opposition to abortion in the Victorian era was from Catholics. The Catholic Church had condemned abortion since its inception("Respect for Unborn"). In contrast, most American Protestant denominations did not have a strong stance against abortion(Peterson). This was partially due to the Catholic Church's limited influence on Protestant Americans(Oakley, et al). As a result, Catholic lobbying had little effect on its own.

These attitudes toward abortion would stay relatively the same until societal norms shifted with the sexual revolution and the women's liberation movements beginning in the 1960s. During this time, the general public began to shift away from the conservative ideals of the Victorian era that had been accepted for the past century(Peterson). For the first time in American history, reproductive rights were starting to be perceived as a women's rights issue(Peterson). Many feminists saw abortion as an extension of bodily autonomy and therefore a human right. However, this opinion was not shared among all feminists. When the National Organization for Women (NOW) "called for the repeal of all abortion laws", many of its

members left the group(Burkett). Some of those feminists later joined the Feminists for Life, an anti-abortion feminist organization still prominent today(Burkett).

The first opponents to NOW and other advocates for reproductive rights, along with the Feminists for Life, were religious individuals and organizations, namely Catholic healthcare professionals, lawyers, and housewives(Holland). One of the most prominent organizations formed by these demographics was the National Right to Life Committee("Abortion"). Today, it asserts itself to be "the nation's oldest and largest grassroots pro-life organization"("About NRCL").

Linking abortion with feminism, however, also linked the anti-abortion movement with anti-feminists(Du Mez). This shift pushed conservative evangelicals who supported segregation and opposed gay marriage to join forces with otherwise less conservative anti-abortionists(Du Mez). Prior to this, American Protestants, including evangelicals, who today are some of the most vocal anti-abortionists, were relatively open-minded towards abortion(Peterson). Many primarily evangelical anti-abortion groups were founded around this time and would only grow larger after "Roe v. Wade" in 1973(Lindsey, "*A Concise History*").

As these social changes occurred, the landscape of abortion providers in America shifted as a result of the changing demographics of the anti-abortion movement from medical professionals to the politically conservative and the religious. Prior to the Victorian era, maternal health was almost completely ignored by physicians(Blakemore). Women would go to a midwife for almost any physical aid, in part because of the commonly held belief that any ailment was due to the menstrual cycle(Floyd). Midwives were almost entirely women, so men had no control over women's health(Blakemore). In addition, professional doctors were rare and only started to become common in the 19th century("Apothecaries"). In 1847, the American Medical Association, or AMA, was founded("Abortion"). However, in the mid 1900s, there were many prevalent alternatives to the modern doctors we think of today, such as midwives, and other so called "quacks"(Annalies). Physicians had relatively little authority, and the AMA needed to assert its dominance over its competition(Holland). An important factor for this is that many prominent AMA members were morally against abortion(Peterson). For example, Horatio Storer, a notorious anti-abortionist gynecologist, pushed the AMA to consider abortion a crime and founded the Physicians Campaign Against Abortion(Blakemore).

All the AMA had to do to achieve their goal was convince American society that physicians were the right choice to monitor pregnancy and childbirth, therefore gaining the authority it needed. There were two ways they achieved this. First, they needed to persuade the public that doctors were an authority on morality as well as medicine. Then, they needed to discredit the "quacks"(Annalies) and leave physicians as pregnant women's only option. The increasingly controversial issue of abortion was used to do both of these things.

In 1857, the AMA began to write letters to state legislators in order to convince them to ban abortion(Peterson). They said that physicians believed abortion was wrong and that it violated the Hippocratic Oath to "do no harm."("Roe v. Wade") The AMA made their stance public, leading religious anti-abortionists to become more supportive of physicians(Peterson).

They began to trust doctors as their values seemingly mirrored their own. In addition, physicians began to undermine women and their ability to make decisions for their own bodies(Holland). Previously, when quickening was used to determine pregnancy, few illegal abortions could be prosecuted(Holland). This weakened doctors' credibility in cases of abortion. In order to gain control, they began to spread the idea that "bodily processes could 'speak for themselves,' though they did need doctors to translate"(Holland). Women were painted as creatures who could not understand their own bodies or be trusted to make decisions, such as having an abortion(Peterson). These ideas drove the public even further away from abortion rights, and they saw doctors as champions for their own morals(Peterson).

In the past, women had been trusted and expected to manage their own health and bodies, but male doctors increasingly gained control. Over the course of the century, abortion would be criminalized in every state, mostly due to the AMA's efforts(Peterson). In order to dissuade people from their competitions, physicians used the widespread animosity towards abortion to sway public opinion away from irregular practitioners such as midwives(Annalies). The AMA and its supporters began to paint them as primarily providers of abortion(Annalies). Midwives were villainized for providing abortions, and public opinion of them turned(Lindsey, "*What Did the Suffragists*"). For example, Madame Restell, publicly called the "Wickedest Woman in New York," was a provider of surgical abortions and abortifacients(Lindsey, "*What Did the Suffragists*"). She was continually berated for providing abortions and was put in jail; ultimately, she committed suicide in 1878.

The AMA had succeeded in becoming the gatekeeper of American medicine(Lindsey, "*What Did the Suffragists*") and physician-led anti-abortion efforts slowed. By the early 20th century, some doctors began to support abortion. During the Great Depression, many surgical abortions were done by physicians. It is likely that this is due to physicians commiserating with women who were reluctant to bring up a child in the current economic conditions. While this greatly slowed down after the second World War, doctors continued to sympathize with women seeking abortions(Annalies). In the 1960s and 1970s, doctors were some of the most prominent advocates for abortion(Peterson). Physicians, once at the forefront of criminalizing abortion, now openly supported reproductive rights and liberalizing abortion laws.

While the religious anti-abortion movement had always influenced the movement, it became the primary catalyst for the continued criminalization and subsequent re-criminalization as the modern pro-abortion movement gained popularity(Holland). This is in part due to the lessening presence of medical professionals in the movement, but mainly due to increasing prevalence of religious anti-abortionists. Catholics remained prominent in the movement, but Protestant anti-abortionists grew exponentially(Du Mez). This had less to do with Christian beliefs towards abortion and more so to do with the rise of the Republican "moral majority."(Du Mez) As mentioned earlier, abortion became lumped in with many other conservative issues to promote Republican politicians. However, many previously neutral denominations began to denounce abortion to reflect the majority of their members' views, namely the Southern Baptist

Convention, the Evangelical Church, and the Methodist Church(Liu). This led to modern anti-abortionists using their religious beliefs to explain their opposition today.

Finally, the anti-abortion movement was greatly changed by developments in the medical fields of gynecology and obstetrics. As previously explained, abortion was only truly considered an abortion if it occurred after "quickening" prior to the Victorian era. While there were some extreme anti-abortionists who viewed abortions before this as immoral, they were generally accepted. However, this gradually changed as scientific discoveries regarding fetal development happened.

One of the ways the AMA convinced legislators to criminalize abortion, especially surgical abortions, was by asserting their danger("Abortion"). While this was true, going through a pregnancy and giving birth were also extremely dangerous at this time(Blakemore). Puerperal fever, for example, often seemingly mysteriously appeared after a woman gave birth and was often fatal(Floyd). In the 1830s, the average woman had more than six children, and about 0.5% of births resulted in death(Withycombe). This may seem low, but since then, the survival rate of giving birth has increased by over 99%(Hoyert). In addition, medicinal abortions were very rarely fatal(Holland).

In the 1850s, the AMA began to assert that life began at conception rather than quickening(Peterson). This was partially due to scientific developments but mostly aided in their campaign to discredit midwives(Holland). This was corroborated by significant advances in the field of embryology in the 1800s by scientists like Karl Ernst von Baer(Oppenheimer). By the early 20th century, more conclusive evidence regarding fetal development supporting this claim began to be discovered("Abortion"). This idea gradually became accepted among the general public. In 1869, the Catholic Church changed their teachings to assert that life began at conception("Abortion"). Any abortion after conception was now considered religiously wrong, instead of after the quickening. This led Catholic anti-abortionists to alter their own beliefs. Almost all early abortion laws regarded regulating abortifacients and post-quickening abortions(Blakemore). However, the latter was difficult to prove as it relied on the pregnant woman's testimony. With the increasingly accepted belief that life began at conception came the idea that aborting a fetus at any point after conception was an abortion.

By the early 20th century, pregnant women now had almost no legal control over their pregnant bodies after several new laws, such as the Comstock Act of 1873 and the Pure Food and Drug Act in 1906, made it virtually impossible to access medicinal abortifacients for any stage of pregnancy(Blakemore). This left desperate women with only one option: illegal, surgical abortions(Peterson). Anti-abortionists made sure that anyone found to have aborted a fetus would be shamed publicly(Ziegler). Despite this, many women still decided on a surgical abortion(Blakemore). It is estimated that by the 1930s, over a million abortions were being performed by sympathetic doctors(Lindsey, "*A Concise History*"). However, surgical abortions were still extremely dangerous, and it was not uncommon for them to end in death(Peterson). Making it harder to access safe, medicinal abortions through herbs had in fact led women to seek dangerous, surgical abortions. The AMA and other anti-abortionists who had once claimed to

oppose abortion due to its danger ignored the fact that its continued criminalization had ultimately caused greater death.

Eventually, the awareness of sepsis and the invention of more effective antibiotics led to a higher success rate in all surgery, including abortions(“Abortion”). In the mid 20th century, surgical abortions were safer than going through with a pregnancy in many cases, although they were unregulated and back-alley abortions done by unqualified practitioners were common(Peterson). It is estimated that around 5,000 people died every year from failed abortions at this time(Dine). After Roe v. Wade, this statistic dropped significantly(Dine). This did not affect anti-abortionists' vigor in their opposition to the issue. Although the origin of the anti-abortion movement was backed by the scientific community, it became more distant from medical fact as the movement and science developed.

To conclude, throughout its history, the anti-abortion movement has changed dramatically. In colonial America, abortion was largely ignored, but in the 19th century it became more relevant. It slowly faded from public conversations until the 1960s when it was once again brought into public consciousness by the cultural revolution. Many different groups have influenced and been a part of the American anti-abortion movement. The American Medical Association used the issue of abortion to reshape the role of physicians in Victorian society. As their efforts turned successful, the medical community began to disregard its previous disapproving stance on abortion and began to become more divided. At the same time, scientific advancements in medicine, such as new understandings of sepsis and fetal development, shifted the general public’s understanding of abortion. Perhaps most importantly, The common idea of a “quickening” in the 18th and 19th centuries was ignored and conception began to be seen as the start of a fetus’s life. Over time, religious anti-abortionists grew in number and influence, especially in the 1960s and 70s as social movements such as the Women’s Liberation Movement also grew in popularity. This marked a turning point in the abortion debate, as the two sides became increasingly polarized. The division has remained to this day, and it has made the topic of abortion largely disputed.

Today, the anti-abortion movement asserts that they intend to conserve old ideals and keep things the way they have always been. However, if you consider the history of the movement as well as the history of abortion itself, it becomes clear that the modern “pro-life” movement is more accurately a reflection of the late 19th and early 20th centuries. For example, the accepted definition of abortion has shifted drastically. Until relatively recently, abortions in the early stages of pregnancy were completely legal and socially acceptable. Aborting a fetus prior to the quickening was not truly considered an abortion(Peterson). In addition, many current anti-abortionists use their religion to explain their beliefs. However, many Christian denominations, such as the Evangelical Church and the Methodist Church, only changed their own teachings on abortion due to the influence of their anti-abortionist members(Holland). Finally, modern anti-abortionists promote their organizations as “pro-life.” They claim to want to save the lives of the unborn, despite the fact that criminalizing abortions does not decrease the number of abortions, only the number of fatalities due to abortions. The current anti-abortion

movement prioritizes the supposed lives of fetuses over the pregnant person and tries to justify this by using their religious beliefs and what they believe to be the ideals of the past.

Works Cited

"Abortion." *Violence in America*, 1999. *Gale In Context: U.S. History*, link.gale.com/apps/doc/BT2350011002/UHIC?u=nysl_me_convent&sid=bookmark-UHIC&xid=379feced. Accessed 19 July 2023.

"About NRLC." *NRCL*, 2019, www.nrlc.org/about/.

"Apothecaries from the Eighteenth Century Onward: USA · Jars of "Art and Mystery": Pharmacists and Their Tools in the Mid-Nineteenth Century · OnView: Digital Collections & Exhibits." *Collections.countway.harvard.edu*, collections.countway.harvard.edu/onview/exhibits/show/apothecary-jars/eighteenth-century-usa. Accessed 19 Dec. 2023.

Barrett, Kara L. *Digital Commons at Buffalo State Victorian Women and Their Working Roles*. 2013.

Blakemore, Erin. "The Complex Early History of Abortion in the United States." *History*, 11 Apr. 2023, www.nationalgeographic.com/history/article/the-complex-early-history-of-abortion-in-the-united-states?loggedin=true&rnd=1689110989862. Accessed 2 Oct. 2023.

Burkett, Elinor. "Women's Rights Movement." *Encyclopedia Britannica*, 6 Nov. 2020, www.britannica.com/event/womens-movement.

Dine, Ranana. "Scarlet Letters: Getting the History of Abortion and Contraception Right." *Center for American Progress*, 8 Aug. 2013, www.americanprogress.org/article/scarlet-letters-getting-the-history-of-abortion-and-contraception-right/.

Floyd, Barbara. *From Quackery to Bacteriology*. *Utoledo.edu*, University of Toledo Libraries, 1994, www.utoledo.edu/library/canaday/exhibits/quackery/quack4.html. Accessed 10 Aug. 2023.

Holland, Jennifer. "Abolishing Abortion: The History of the Pro-Life Movement in America - Organization of American Historians." *Organization of American Historians*, Organization of American Historians, 23 Aug. 2019, www.oah.org/tah/november-3/abolishing-abortion-the-history-of-the-pro-life-movement-in-america/.

Hoyert, Donna. "Maternal Mortality Rates in the United States, 2021." *Www.cdc.gov*, 16 Mar. 2023, www.cdc.gov/nchs/data/hestat/maternal-mortality/2021/maternal-mortality-rates-2021.htm#:~:text=The%20maternal%20mortality%20rate%20for.

Lindsey, Treva B. "A Concise History of the US Abortion Debate." *The Conversation*, 10 June 2019, theconversation.com/a-concise-history-of-the-us-abortion-debate-118157.

Lindsey, Treva B. "What Did the Suffragists Really Think about Abortion?" *Smithsonian Magazine*, 26 May 2022, www.smithsonianmag.com/history/what-did-the-suffragists-really-think-about-abortion-180980124/. Accessed 12 July 2023.

Liu, Joseph. “Religious Groups’ Official Positions on Abortion.” *Pew Research Center’s Religion & Public Life Project*, 16 Jan. 2013, www.pewresearch.org/religion/2013/01/16/religious-groups-official-positions-on-abortion/.

Martin, Michel, and Kristen Kobes Du Mez. “How Abortion Became a Mobilizing Issue among the Religious Right.” *NPR.org*, NPR, 8 May 2022, www.npr.org/2022/05/08/1097514184/how-abortion-became-a-mobilizing-issue-among-the-religious-right. Accessed 10 Aug. 2023.

Oakley, Francis Christopher, et al. “Roman Catholicism.” *Encyclopædia Britannica*, www.britannica.com/topic/Roman-Catholicism. Accessed 29 Dec. 2023.

Oppenheimer, Jane M. “Karl Ernst von Baer .” *Encyclopedia Britannica*, 27 Apr. 2017, www.britannica.com/biography/Karl-Ernst-Ritter-von-Baer-Edler-von-Huthorn. Accessed 27 Dec. 2023.

Peterson, Anna M. “From Commonplace to Controversial: The Different Histories of Abortion in Europe and the United States.” *Origins: Current Events in Historical Perspective*, Nov. 2012, origins.osu.edu/article/commonplace-controversial-different-histories-abortion-europe-and-united-states?language_content_entity=en.

“Respect for Unborn Human Life: The Church’s Constant Teaching | USCCB.” *Www.usccb.org*, United States Conference of Catholic Bishops, 2022, www.usccb.org/issues-and-action/human-life-and-dignity/abortion/respect-for-unborn-human-life.

“Roe v. Wade.” *Gale U.S. History Online Collection*, 2023. *Gale In Context: U.S. History*, link.gale.com/apps/doc/NISSCS973716604/UHIC?u=nysl_me_convent&sid=bookmark-UHIC&xid=3348abe1. Accessed 6 July 2023.

Winny, Annalies. “A Brief History of Abortion in the U.S.” *Hopkins Bloomberg Public Health Magazine*, 26 Oct. 2022, magazine.jhsph.edu/2022/brief-history-abortion-us.

Withycombe, Shannon K. “Women and Reproduction in the United States during the 19th Century.” *Oxford University Press*, 25 Jan. 2019, oxfordre.com/americanhistory/display/10.1093/acrefore/9780199329175.001.0001/acrefore-9780199329175-e-426. Accessed 19 Dec. 2023.

Ziegler, Mary. “Some Form of Punishment: Penalizing Women for Abortion.” *William & Mary Bill of Rights Journal*, vol. 26, no. 3, Mar. 2018, scholarship.law.wm.edu/cgi/viewcontent.cgi?article=1851&context=wmborj. Accessed 19 Dec. 2023.

The CEO Formula: Exploring Ownership, Compensation Structure & Demographics as Determinants of Firm Performance By Richard Schäli

Abstract

This paper investigates the relationship between Chief Executive Officers' ownership stakes, compensation structures, and demographic characteristics with firm performance in the S&P 500 index over the period 2014-2024, drawing on quantitative panel data that combines time series and cross-sectional dimensions. The paper uses Agency Theory and Upper Echelons Theory to examine how the aforementioned factors influence key performance indicators, including Return on Assets (ROA), Return on Invested Capital (ROIC), and stock performance. The findings indicate that equity-heavy compensation packages align CEO and shareholder interests, increase operational efficiency and thus improve stock performance. However, excessive reliance on equity-based pay may encourage short-term risk-taking. CEOs with larger ownership stakes tend to achieve higher ROA and ROIC through increased efficiency but only have limited influence on stock market outcomes. Demographic factors like tenure and age correlate with operational stability but can also dampen innovation and growth in dynamic industries. The results show that while compensation and ownership have a moderate impact on operational performance, external market pressures and firm-specific conditions have a much greater impact. This research shows that incentive models are powerful but must be tailored to the specific conditions of a company's industry and circumstances. There is no universal blueprint for executive remuneration, and compensation structures alone cannot ensure sustainable success.

1. A Modern Corporate Puzzle

The role of the Chief Executive Officer (CEO) in the modern corporate landscape cannot be understated. CEOs lead organizations, set strategic vision, foster corporate culture, and drive critical decision-making processes that determine a firm's success or failure. This paper examines what factors enable CEOs to drive superior firm value. One critical area of focus within this discipline is CEO compensation. Compensation packages serve as a tool for mitigating the inherent conflict of interest between managers and shareholders. Properly designed compensation structures can align CEOs' goals with those of the shareholders and motivates executives to prioritize long-term value creation over short-term gains. In this context, CEO compensation has arisen as the key governance mechanism for addressing agency problems and incentivizing performance. Agency problems occur when a CEO's interests conflict with those of shareholders, especially when the CEO has no financial stake in the company's outcomes (Boyd, 1994). By linking a CEO's compensation to stock performance through equity-heavy compensation packages, a board aligns the CEO with shareholders to maximise shareholder value.

Corporate governance mechanisms are designed to influence CEO behaviour and mitigate agency problems. Besides compensation, these mechanisms include ownership

structures and board characteristics such as composition, diversity and the degree of board intervention in the CEO's work. While the board's primary role is to monitor the CEO's discretion and hold him accountable for his decisions, it is equally important not to remove that discretion entirely. As Lilienfeld-Toal and Ruenzi (2013) highlight, when high ownership is combined with significant CEO discretion, it can amplify abnormal stock price returns.

Much of the traditional strategic management literature focuses on processes and decisions (such as resource allocation, strategic planning, and performance monitoring) as abstract constructs (Aguilar, 1967; Allen, 1979). To address this, those studies focused on aligning incentives by structuring compensation, bonuses, and rewards to align managers' interests with shareholder goals. This approach reduces conflicts of interest but neglects the influence of individual traits on decision-making. For the first time, Hambrick and Mason's *Upper Echelons Theory (UET)* challenged this, by emphasizing that executives' personalities, life experiences, and perspectives directly shape strategic choices and organizational outcomes. The UET shifts the focus to *who* the CEO is, bridging the gap between compensation structures and the personal characteristics of top executives, which can partially predict strategic choices and performance levels (Hambrick & Mason, 1984).

The relationship between CEO demographics, ownership, and compensation structures and their influence on firm performance has been the subject of extensive research. This study builds on that established body of work, positioning itself alongside it while addressing the evolving nature of corporate landscapes. As business environments and governance practices continuously change, older findings may no longer fully reflect current market conditions or emerging trends. By utilizing an up-to-date dataset spanning 2014 to 2024 and integrating variables that previous studies have often addressed separately, this paper offers a comprehensive and contemporary perspective. Furthermore, there is a wide diversity of the methodological approaches and definitions of key variables, such as firm performance, used in those studies. There is a necessity for periodic re-examinations to ensure the continued relevance and applicability of governance models. Ultimately, this study aims to clarify how demographic characteristics, ownership stakes, and compensation models collectively shape corporate outcomes, offering practical insights for boards seeking to design effective executive governance structures.

This paper aims to synthesize these disparate strands of research to develop a more holistic understanding of what makes an effective CEO. The central research questions guiding this study include: Does higher compensation correlate positively with firm performance? Does high CEO ownership lead to better alignment with shareholder interests? How do demographic and cultural attributes of CEOs influence strategic decision-making and innovation? This paper focuses specifically on publicly traded companies within the S&P 500 index from 2014 to 2024. This limits the results to U.S. firms, excluding private companies and those outside the U.S., and restricts the analysis to short-term changes rather than long-term trends in corporate governance practices.

2. Theoretical Foundations: Bridging Incentives and Leadership

2.1 Agency Theory

Agency Theory addresses the inherent conflict of interest between shareholders (principals) and CEOs (agents). Shareholders seek to maximize firm value, whereas CEOs, as rational actors, aim to maximize their personal gain (Jensen & Meckling, 1976). This divergence creates the "agency problem," where CEOs may prioritize actions that benefit themselves at the expense of shareholders. One of the basic approaches to deal with this issue is through equity-based compensation, such as stock options or performance shares. By tying the CEO's rewards to the company's performance, it successfully transforms the CEO into a partial principal and aligns their incentives to act according to shareholders' interests (Eisenhardt, 1989). This conflict arises because of information asymmetry, where agents (CEOs) have more knowledge about the firm's operations and strategic direction than the principals. This divergence makes it very difficult for shareholders to monitor and ensure that CEOs act in their interest (Fama & Jensen, 1983). Additionally, the separation between ownership and management exacerbates the issue since CEOs will not always have an incentive to prioritize shareholder value over their self-interests, often leading to decisions that maximize short-term financial gains or personal perks at the expense of long-term organizational health (Tirole, 2001). In order to reduce agency cost, by aligning incentives of agents with those of the principal, boards can put performance-based compensation, such as stock options, restricted shares, and other corporate governance structures in place (Core et al., 1999). Despite such mechanisms, the agency problem is one of the central challenges in corporate management, as complete alignment of interests is near-impossible to fully bring into effect (Shleifer & Vishny, 1997).

However, the principal-agent relationship outlined in Agency Theory relies on several simplifying assumptions that might not be realistic in practice. One of the key assumption that CEOs are rational agents, who maximize their own utility are oversimplifications since humans not always behave in self-interested and predictable ways; behavioral factors such as overconfidence, loss aversion, or short-termism often interfere with decision-making processes (Gabaix & Landier, 2008). A notable real-life example is Adam Neumann, former CEO of WeWork, whose overconfidence played a significant role in the company's near-collapse. At its peak, WeWork was valued at $47 billion, driven by Neumann's aggressive growth strategy and visionary rhetoric. He pursued rapid global expansion without a clear path to profitability, overestimating market demand for shared office spaces. The company's 2019 IPO filing revealed $1.6 billion in losses on $1.8 billion in revenue the prior year and raised concerns over Neumann's leadership. Additionally, the filing highlighted questionable practices, such as self-dealing through leasing properties he personally owned to WeWork. Following the IPO's collapse and the company's valuation dropping to $8 billion, shareholders suffered massive losses, ultimately leading to Neumann's ousting from the company (Robertson, 2019).

2.2 Upper Echelons Theory

Compared to Agency Theory, the more modern Upper Echelons Theory (UET) argues that organisational performance is driven by the personal characteristics of managers; cognitive bases, values, education, experience, tenure and soft skills. Such characteristics influence how managers perceive incentives, so that risk-takers and risk-averse individuals will respond differently to incentives, which in turn affects strategic decisions and firm performance. While agency theory argues that firm outcomes are driven by exogenous factors such as equity compensation, UET clarifies the relationship between four important elements: strategic contexts, top-level characteristics, strategic decision-making, and firm outcomes (Hambrick & Mason, 1984).

Managers' cognitive traits act as a "perceptual screen" (Hambrick & Mason, 1984, p. 195) for interpreting stimuli and making strategic choices. This screen operates through three mechanisms: limited vision, where biases narrow focus on certain stimuli; selective perception, which limits information processing; and interpretation, where cognitive frames influence the understanding of strategic contexts (Hambrick, 2018). Observable demographic characteristics serve as proxies for the cognitive bases and values of executives and offer strong predictors of strategic decisions and firm performance (Carpenter et al. 2004). For instance, age, education, functional background, and tenure all mold decision-making. Younger managers are more likely to take risks and innovate; older managers are risk-averse and conservative in their desire to preserve the status quo (Wiersema & Bantel, 1992). Similarly, educational background is a proxy for cognitive ability and tolerance for ambiguity, such that more highly educated managers pursue more innovative and strategic changes (Bantel & Jackson, 1989). Functional background also affects the decision-making processes; managers in output functions, such as marketing, for instance, prefer growth strategies, whereas those in throughput functions - production, for example - emphasize efficiency (Geletkanycz & Black, 2001).

Hambrick & Mason (1984) also urge to examine complete top management teams, TMTs, and not just chief executive officers. The diversity within TMTs, which represents different cognitive bases, leads to better decision quality, increased creativity, and greater strategic flexibility - especially in times of crisis (Nielsen, 2009). Too much diversity, however, can cause conflict and lessen cohesion, thus weakening the soundness of decision-making (Hambrick et al., 1996).

2.3 Compensation Structures & Corporate Performance

Empirical studies investigate the link between executive compensation structures and firm performance. In this secondary-literature synthesis several important implications regarding compensation systems can be found. First, the form of compensation - cash, equity, or stock options - has a more significant effect on managerial incentives and organizational outcomes than does the level of compensation. Second, while cash compensation may provide security and decrease individual financial risk, it often produces an alignment gap between managerial interests and shareholder concerns. In contrast, equity-based and option-based compensation schemes remedy this gap by directly relating compensation to firm performance. Third, effective

governance would ensure compensation structures are designed to encourage managers to take appropriate risks while avoiding excessive risk-taking.

2.3.1 Salary Cash Compensation

Cash salaries are predictable, they do not motivate the managers to maximize firm performance because their payouts are not tied to the stock performance or any other indicators. A high reliance on fixed cash salaries often results in a disconnect, as managers are not directly rewarded for increasing stock performance or achieving key performance indicators (Jensen & Murphy, 1990). Empirical studies support this, in that managers - being risk-averse - will always prefer fixed cash compensation rather than performance-related pay, mainly to reduce their personal financial risk (Harris and Raviv 1979).

2.3.2 Equity-Based Compensation

Managers who own greater equity stakes are motivated to prioritize the long-term development of the firm, thereby mitigating agency conflicts (Jensen & Meckling, 1976). Empirical studies demonstrate a clear positive correlation between return on assets (as a proxy for firm performance), and the percentage of compensation tied to equity (Mehran, 1995). Another important finding by Mehran is that compensation structure, not the level of pay, is what drives managerial efforts toward maximizing corporate efficiency.

2.3.3 Stock Options

Stock options provide a conditional form of equity-based compensation and grant managers the right to purchase shares at a predetermined price in order to elicit future performance improvements. Excessive reliance on stock options may encourage short-term stock price manipulation or riskier-than-optimal strategies. For example, Enron executives were awarded millions in stock options, with CEO Jeffrey Skilling earning $132 million in 2000 alone. This heavy reliance on options incentivized executives to manipulate earnings reports and inflate stock prices by over 500% from 1998 to mid-2000, before the company's collapse in 2001 (Securities and Exchange Commission, 2003) Poorly designed incentives, whether for executives or directors, can undermine governance and harm long-term performance. Without significant equity ownership, outside directors lack a direct financial stake in the organization's success. This limits their motivation to actively monitor management's decisions and actions since their personal compensation is not critically tied to the firm's performance outcomes (Crystal, 1991).

2.3.4 Ownership

Ownership interacts with compensation structures by naturally aligning the interests of CEOs and shareholders. For example, owner-CEO companies might rely less on salary-based pay because the CEO's equity stake already incentivizes long-term value creation. Empirical studies illustrate the impact alignment between agent and principal can have. Lilienfeld-Toal and Ruenzi (2013) conducted an influential study published in the *Journal of Finance*, which found

that investing in the top 10th percentile of "owner-operator" companies - where CEOs hold a significant degree of voluntary equity ownership - generated annual abnormal returns of 5% from 1980 to 2010 across all U.S. equities. The study excluded options and restricted shares to focus on ownership that CEOs chose voluntarily. When CEOs buy shares with their own money, rather than receiving them as part of their compensation, it shows a personal commitment to the company. This commitment is likely driven by the expectation of earning higher returns, despite the risks of holding a less diversified portfolio.

Those above-market returns examined by Lilienfeld-Toal and Ruenzi are based on the critical agency issue of information asymmetry. CEOs, as insiders, possess superior knowledge about the company's potential performance. When they hold significant stock, markets often interpret this as confidence based on insider insight, thus expecting above-market returns. (Lin & Howe, 1990). However, such ownership may also be a signaling mechanism aimed at projecting strength to investors, potentially boosting short-term firm value without sustaining long-term performance (Leland & Pyle, 1977). Another explanation for those abnormal returns in owner-CEO companies, lies in CEO stock ownership as a corporate governance device (Gompers et al. 2003; Cremers & Ferrell 2009). Equity alignment reduces agency costs by incentivizing CEOs to focus on efficiency. Owner-operators often accept lower compensation, relying on firm performance for their personal wealth. This alignment drives cost control, higher productivity, and lower overheads, reflected in superior returns on assets. This governance dynamic creates long-term value, making this explanation particularly compelling.

2.4 Association Between Demographics & Corporate Performance

2.4.1 Manager Discretion

Homogeneous top management teams make strategic decisions faster, which benefits stable environments, while heterogeneous teams outperform in turbulent or discontinuous settings (Hambrick & Mason, 1984). Lillienfeld-Toal & Ruenzi (2011) demonstrate that firms with high managerial discretion-which they define as having considerable CEO power through low institutional ownership, or a founder-status, realize higher abnormal returns when combined with strong ownership incentives.

Greater managerial discretion improves the effectiveness of particular compensation schemes through the ability of CEOs to exercise their power and resources in implementing strategic actions for better outcomes under the right incentives. Without adequate discretion, optimal pay incentives are not able to improve organizational performance because CEOs feel they lack the ability to influence key decision-making processes (Lillienfeld-Toal & Ruenzi, 2011).

2.4.2 Education

A CEO's higher education is often the last formal training that polishes their personality and skills.

Although this refers to a different geographical context, a study examined 3,902 firm-year observations of FTSE 350 companies from 1999 to 2017. It used industry-adjusted Return on Assets (ROA) as a proxy for firm performance. The study, published by Urquhart and Zhang (2022), found that CEOs with undergraduate, postgraduate, or MBA degrees had no significant impact on firm performance. The coefficients for these degrees were statistically insignificant. In contrast, CEOs with PhDs demonstrated a more significant performance benefit, increasing industry-adjusted ROA by 3.03%.PhDs from the top 100 universities delivered an even greater impact, boosting ROA by 4.65% - a 1.62% premium over non-elite PhDs. Results in the study are all consistent that PhD-level education, especially from elite schools, improves firm performance, while other qualifications have little impact (Urquhart & Zhang, 2022).

2.4.3 Tenure

Some papers have found an inverted U-shaped relationship between CEO tenure and the firm performance, where performance would improve up to a certain point of tenure before declining (Guang-sheng, 2010). This pattern is consistent with findings on acquisition profitability and, in general, consistent with learnings in other contexts, where CEOs become more effective in their early tenure as they learn the specific job and adapt to it. In the later stages of their tenure, acquisition effectiveness decreases due to entrenchment, broadened power bases, and strategic inflexibility, resulting in the subsequent lower shareholder returns (Walters et al. 2007). At the same time, as CEOs' tenure proceeds, inside equity ownership rises for most and better aligns the interests with those of shareholders; it impels stronger incentives toward better firm value performance. The changing pattern of ownership will, therefore, offset entrenchment by motivating steady improvement in the firm's value.

2.4.4 Cultural/Socioeconomic Background

Horton et al. (2012) argue that no theoretical framework suggests that nationality affects performance, but this view may overlook the dynamic interaction between a CEO's cultural fit and their organisation. Nationality may serve as a proxy for global experience or leadership style, and incongruence between a CEO's background and organisational culture may hinder performance (Sebbas, 2017). On the other hand, companies with foreign CEOs often surpass all expectations, perhaps because of the different viewpoints and specialized talents that these individuals bring with them (Kaufmann, 2004). The cultural diversity of top management teams can be a source of both challenges and opportunities. Drawing on social identity theory, Turner (1982) suggests that non-task-related diversity, such as cultural differences, can strain team cohesion, while shared backgrounds can facilitate mentoring and information exchange (McDonald & Westphal, 2013).

2.4.5. Personality Traits and Leadership styles

The Big Five personality framework identifies Conscientiousness, Extraversion, Openness, Agreeableness and Emotional Stability as some of the most important traits that

influence CEO behaviour and performance (Kurdyukov, 2023; Judge & Piccolo, 2004). For example, conscientious CEOs are very good at setting clear goals and fostering a culture of accountability; extraverts are charismatic and inspire people. However, this natural confidence can sometimes turn into an overconfident CEO who neglects critical feedback, underestimates risks, or overstretch resources in ways that lead to either a flawed strategy or inefficiencies in the organisation's approach (Malmendier & Tate, 2005). Openness leads to innovation and adaptability and helps develop things that provide a competitive advantage and ensures for a company to survive cycles of technological disruption and agreeableness promotes collaboration and conflict resolution within an organisation (Araujo-Cabrera et al., 2017). Emotional stability can provide CEOs with the skills needed to manage stress and crises competently, but at the same time it can promote certain bureaucratic tendencies (Ormiston et al., 2022).

2.5. Research Gaps in Strategic Leadership Studies

Following the literary review there is an important research gap to be found, which is to identify the specific types of equity-based compensation that best align the interests of management and shareholders. While existing secondary literature has shown that ownership in equity, often gained through the accrual of equity compensation, generally aligns the interests of agents with those of principals, there remains a large gap in understanding the contextual application of the specific types and structures of equity-based compensation that are best suited to achieving these results. The combination of implications from Upper Echelons Theory explains that the interaction between compensation structures and managerial behavior is moderated by context variables such as a manager’s characteristics and the existing culture within the organization. Interestingly, these contextual moderators are relatively poorly explored within empirical research. This gap in existing research can be attributed to the rarity of adopting more holistic case studies that examine how equity-based compensation is practiced. Such studies could shed light on how compensation structures affect managerial soft skills and decision-making, with the goal of developing practical principles that would help financial professionals make wiser decisions regarding executive compensation. To address this research gap, this paper designed a quantitative methodology that examines the relationship between compensation, manager’s age and tenure, and firm performance . By leveraging financial and demographic data from S&P 500 firms, the study aims to provide empirical insights into how these factors interact in practices.

3. Methodology

3.1 Prologue

This paper aims to investigate efficient governance structures to incentivize CEOs and align their actions with shareholder interests. By building on the existing body of knowledge, this paper should further investigate the interaction between equity-based compensation and ownership stakes with CEO demographic characteristics (such as age and tenure) in order to

identify the unique characteristics that make great CEOs“. More specifically, it should be deduced whether higher levels of equity stakes, certain equity-to-cash ratios in compensation packages, and a particular age or tenure of CEOs result in better corporate outcomes. Answering these questions calls for a quantitative approach of research by using financial and demographic data for the companies included in the S&P 500 index, for the period 2013–2023. However, it should be acknowledged that there are certain limitations to this study. For example, data on ownership do not distinguish between voluntary and involuntary holding, which could skew the interpretation of the results. Also, this paper will not adjust stock performance for risk factors using models like the Carhart four-factor model. While other datasets include these risk factors to address inconsistencies in stock returns by accounting for the level of risk associated with certain returns, this study does not incorporate them due to their complexity and the high-level scope of the research. Given the limited scope of the dataset, the analysis focuses on US publicly traded firms within a very narrow timeframe, which makes the paper's findings non-generalizable to private firms, markets outside of the United States, as well as to findings in different contexts of time.

3.2 Research Design and Approach

This study builds upon Agency Theory and Upper Echelons Theory to explore how CEO characteristics and compensation structures influence firm performance. Specifically, the research focuses on the relationship between independent variables such as equity-to-cash compensation ratio, CEO ownership share, age, and tenure, and the dependent variable of firm performance. Firm performance is measured using stock performance, return on invested capital (ROIC), and return on assets (ROA).

The equity-to-cash compensation ratio is defined as the proportion of equity-based compensation, including stock options and restricted stock, to cash-based compensation for each CEO. CEO ownership share refers to the percentage of total equity held by the CEO, encompassing stock options and common shares. Demographic variables such as age are measured at the end of the fiscal year, while tenure is calculated as the total number of consecutive years served in the CEO role at the same firm.

Firm performance metrics include stock performance, measured as the average annual stock return; ROIC, calculated as Net Operating Profit After Taxes divided by Invested Capital; and ROA, determined by dividing Net Income by Total Assets.

To examine these relationships, the study hypothesizes the following. One would expect that higher executive compensation aligns the decision-makers incentives with shareholders, enhancing firm performance. Therefore, the hypotheses are:

H_0: There is no significant relationship between the executive amount of compensation and firm performance.
H_1: Higher executive compensation is positively associated with firm performance.

Similarly, it could be assumed that greater CEO ownership stakes align their interests with those of the shareholders, potentially maximizing returns. Thus, the hypotheses are:

$\mathbf{H}_0$: CEO ownership stake does not significantly affect firm performance.
$\mathbf{H}_1$: A larger CEO ownership stake positively correlates with firm performance.

Finally, demographic characteristics such as age and tenure are included as proxies for experience and organizational stability. Based on Upper Echelons Theory, it is hypothesized that these characteristics influence firm performance:

$\mathbf{H}_0$: CEO age and tenure have no significant effect on firm performance.
$\mathbf{H}_1$: CEO age and tenure are significantly associated with firm performance.

Data for this analysis are sourced from Bloomberg Terminal, which provides detailed breakdowns of compensation, ownership, and demographic data in a set of 504 companies. Compensation data include equity- and cash-based components, while ownership data detail the CEO's holdings of equity (voluntary and involuntarily through options). Demographic data points, including age and tenure, are recorded at the time of observation. By leveraging these datasets, the study aims to contribute to the literature on CEO effectiveness and firm performance, guided by both Agency Theory and Upper Echelons Theory.

3.3 Statistical Analysis

In this study correlation analysis and multiple regression is to be used to evaluate the relationships between CEO characteristics, compensation structures, and firm performance. Correlation identifies relationships between variables but does not establish causation. Regression, on the other hand, isolates specific effects by controlling for confounding factors such as industry, firm size, and market conditions.

This paper employs a simple linear regression model without control variables, and future research is encouraged to include these factors for a more comprehensive analysis. During data preparation, missing or invalid data points are excluded to ensure analytical integrity. To test the hypotheses, Pearson correlation coefficients are computed to measure the strength and direction of bivariate relationships. For example, CEO ownership, age, tenure, and compensation are analyzed in relation to stock returns, ROIC, and ROA, before incorporating these variables into a regression model.

3.4 Model

The analysis begins with data preprocessing to ensure consistency and completeness. Missing or incomplete data points are excluded, and variables are standardized to allow comparability across firms and time periods. As part of the preliminary analysis, descriptive statistics are generated to summarize key features of the dataset and compute Pearson correlation

coefficients to identify potential relationships between variables, ensuring no significant multicollinearity exists (correlations between independent variables remain below 0.7). Following this, multiple linear regression (MLR) models are constructed to examine the relationships between the dependent variables (ROA, ROIC, and stock performance) and the independent variables (CEO tenure and age). Sample equations include:

$$Firm\ Performance\ =\ Intercept \pm X1\ *\ Avg.\ Compensation\ \pm X2\ *\ Compensation\ Growth$$
$$Firm\ Performance\ =\ Intercept \pm X1\ *\ CEO\ Ownership \pm X2\ *\ Inside\ Ownership\ (ex.\ CEO)$$
$$Firm\ Performance\ =\ Intercept \pm X1\ *\ Tenure \pm X2\ *\ Age$$

Using Excel's Analysis ToolPak, regression analysis is performed, generating outputs such as regression coefficients, p-values, and R^2 values to evaluate the strength and significance of these relationships. Correlation coefficients are interpreted using standard thresholds, where values between 0.7 and 1.0 (or -0.7 and -1.0) indicate strong correlations, 0.3 to 0.7 (or -0.3 to -0.7) represent moderate correlations, and below 0.3 (or -0.3) are considered weak. Although multicollinearity has not been explicitly tested, the correlation table confirms that this issue is unlikely in the dataset. Finally, the regression outputs are interpreted in the discussion section, connecting the findings to existing literature and translating them into actionable recommendations for practices.

4. Results & Discussion

4.1 Impact of Executive Compensation on Firm Performance

This analysis shows a significant positive relationship between executive compensation and firm performance, rejecting the null hypothesis (H_0) and supporting the alternative hypothesis (H_1). In straightforward terms, the data indicates that how much executives are paid—and how their pay changes over time—has a measurable impact on how well a company's stock performs. However, it is also important to note that causality may run in the opposite direction, with strong stock performance leading to higher executive pay. The regression model, summarized in Table A1, reveals an F-statistic of 11.81 ($p = 0.00001$) and explains about 4.94% of the variation in stock returns ($R^2 = 0.0494$). Although the model's explanatory power is modest, it confirms that compensation practices play an important role in shaping stock performance.

$$Total\ Stock\ Return\ =\ 125.05\ +\ 1.89\ *\ Avg.\ Comp.\ +\ 0.81\ *\ Comp.\ Growth$$

Two findings stand out. First, the average executive compensation from FY16 to FY24 (Fiscal Year 2016 to Fiscal Year 2024) has a clear positive impact on stock returns. For every $1 million increase in average executive pay, total stock returns increase by 1.89 percentage points (coefficient = 1.89, $p = 0.00776$, Table A2). Second, the growth in executive compensation over time is equally impactful. A 1% increase in compensation growth (measured as the % change from FY16–FY17 to FY23–FY24) correlates with a 0.81 percentage point increase in stock

returns (coefficient = 0.81, p = 0.00004, Table A2). These insights underscore that both the level and the growth of executive compensation are essential drivers of performance.

Table A1: Regression Model Summary – Executive Compensation and Stock Returns Model fit, significance levels, and key metrics for executive pay and stock performance analysis.

Metric	R^2	Adj. R^2	F-statistic	Significance F	Standard Error	Observations
Total Stock Return	0.0494	0.0452	11.81	0.00001	446.53	457

Table A2: Regression Coefficients – Executive Compensation and Stock Returns
Regression results linking executive compensation and growth to stock returns.

Metric	Variable	Coefficient	Standard Error	t-Statistic	P-value	95% Confidence Interval
Total Stock Return	Intercept	125.05	37.65	3.32	0.00097	[51.05, 199.04]
Total Stock Return	Avg Compensation in Millions(FY16-FY24)	1.89	0.71	2.67	0.00776	[0.50, 3.28]
Total Stock Return	% Comp Growth (16-17 to 23-24)	0.81	0.19	4.16	0.00004	[0.43, 1.19]

The correlation between total stock return and average compensation (FY16–FY24) is 0.12, and the correlation with compensation growth is 0.18, indicating a positive but moderate relationship. This analysis highlights the importance of compensation practices in aligning executive performance with shareholder interests.

However, the structure of compensation may be even more critical than the amount. As noted in research by Murphy (1999) and Core et al. (1999), performance-linked pay structures, such as stock options or equity-based rewards, are more effective at driving long-term shareholder value than increases in fixed salaries. Poorly designed compensation packages can incentivize executives to focus on short-term financial engineering, like inflating free cash flow, rather than fostering sustainable value creation. This risk underscores the importance of designing pay structures that promote ethical and effective governance.

Looking at the descriptive statistics from Table A1, we see considerable variability in total stock returns, with an average of 125.05 and a standard error of 37.65, reflecting the diverse performance dynamics across firms in the sample. The modest correlation of 0.12 between stock returns and average compensation suggests that while compensation levels matter, their impact is only part of the story. Meanwhile, the stronger correlation of 0.18 between stock returns and compensation growth reinforces the idea that dynamic compensation strategies are particularly influential in driving better outcomes.

Stock-based compensation, while effective for aligning incentives, can also create risks. Companies that rely heavily on stock-based rewards may inadvertently encourage executives to

prioritize short-term stock price gains over sustainable long-term growth. These concerns highlight the need for rigorous oversight of compensation structures to balance incentive alignment with corporate sustainability.

Lastly, the modest explanatory power of the regression model ($R^2 = 0.0494$, Table A1) reminds us that executive compensation is only one piece of the puzzle. Strong leadership cannot fully overcome weak business fundamentals. Companies with poor competitive positioning, limited market opportunities, or operational inefficiencies are unlikely to excel regardless of executive pay. Conversely, businesses with strong economic moats and favorable industry conditions can deliver exceptional results, even with average leadership.

In conclusion, this analysis confirms that both the level and growth of executive compensation significantly influence stock performance. However, the structure of compensation plays an equally important role in ensuring long-term value creation. These findings emphasize the need for firms to adopt well-designed, performance-linked compensation policies. In conclusion, this analysis confirms that both the level and growth of executive compensation significantly influence stock performance. However, the structure of compensation plays an equally important role in ensuring long-term value creation. These findings emphasize the need for firms to adopt well-designed, performance-linked compensation policies. The relatively low R^2 value may suggest that factors beyond executive compensation - such as strong business fundamentals - are more critical to firm performance.

4.2 Impact of Ownership on Firm Performance

The analysis rejects the null hypothesis that CEO ownership stake does not significantly affect firm performance and supports the alternative hypothesis that a larger ownership stake positively correlates with performance metrics like ROA and ROIC. However, no significant relationship is found between CEO ownership and stock performance, suggesting that while ownership may drive operational efficiency, it does not directly influence market-based outcomes. This conclusion is supported by the regression results in Table A3, which show an R^2 of 0.024 for ROIC and 0.015 for ROA, both statistically significant at $p = 0.0029$ and $p = 0.0246$, respectively. These small R^2 values are typical in multifactor, large-dataset analyses, consistent with Cohen's (1988) observation that small effect sizes are often the norm in the social and behavioral sciences. Conversely, the R^2 for stock performance is only 0.001 with a p-value of 0.733, indicating no reliable relationship.

$$Stock\ Performance = 341.64 + 26.89 * (\%CEO\ Ownership) + 0.74 * (\%Other\ Insiders)$$
$$ROIC = 11.67 + 1.31 * (\%CEO\ Ownership) + 0.003 * (\%Other\ Insiders)$$
$$ROA = 7.32 + 0.59 * (\%CEO\ Ownership) - 0.031 * (\%Other\ Insiders)$$

A 1% increase in CEO ownership is associated with a 1.31 percentage point improvement in ROIC ($p = 0.0007$) and a 0.59 percentage point increase in ROA ($p = 0.0085$), as shown in Table A4. These findings suggest that CEOs with higher ownership stakes tend to manage resources more efficiently, likely because their interests are better aligned with shareholders.

However, insider ownership (excluding CEOs) has no significant effect on ROIC (p = 0.986) or ROA (p = 0.729). Similarly, neither CEO ownership (p = 0.432) nor insider ownership (p = 0.957) impacts stock performance, as outlined in Table A5. This may indicate that external factors, such as market conditions and investor sentiment, overshadow the influence of internal ownership structures.

Table A3: Regression Model Summary – Impact of Ownership on Firm Performance
Summary of ownership's impact across performance metrics.

Metric	R^2	Adjusted R^2	F-statistic	Significance F	Standard Error	Observations
ROIC	0.024	0.02	5.92	0.0029	13.75	486
ROA	0.015	0.011	3.73	0.0246	8.07	488
Stock Performance	0.001	-0.003	0.31	0.733	1224.24	456

Table A4: Regression Coefficients – Impact of Ownership on Firm Performance
Coefficients and significance of CEO and insider ownership on ROIC, ROA, and stock performance.

Metric	Variable	Coefficient	Standard Error	t-Statistic	P-value	95% Confidence Interval
ROIC	**Intercept**	11.67	0.68	17.15	<0.0001	[10.33, 13.00]
ROIC	**% CEO Ownership**	1.31	0.38	3.41	0.0007	[0.56, 2.06]
ROIC	**% Other Insiders**	0.003	0.15	0.02	0.986	[-0.30, 0.30]
ROA	**Intercept**	7.32	0.4	18.37	<0.0001	[6.54, 8.10]
ROA	**% CEO Ownership**	0.59	0.22	2.64	0.0085	[0.15, 1.04]
ROA	**% Other Insiders**	-0.031	0.089	-0.35	0.729	[-0.21, 0.14]
Stock Performance	**Intercept**	341.64	62.58	5.46	<0.0001	[218.66, 464.62]
Stock Performance	**% CEO Ownership**	26.89	34.16	0.79	0.432	[-40.25, 94.03]
Stock Performance	**% Other Insiders**	0.74	13.65	0.05	0.957	[-26.08, 27.56]

These findings differ from prior research, such as Ruenzi and Lilienfeld-Toal (2013), which documented a significant positive relationship between high CEO ownership and abnormal stock returns, using a different year span, methodology, and performance metric (abnormal returns). Their work attributes these returns to strong CEO incentives not immediately priced into the market, suggesting a lag in investor recognition. While this study finds no direct link between CEO ownership and stock performance, it aligns with Ruenzi and Lilienfeld-Toal in observing that high ownership correlates with slower growth and reduced M&A activity.

The regression analysis shows that ownership variables, such as % CEO ownership and insider ownership, as well as leadership characteristics like CEO tenure and age, have minimal impact on revenue growth. The R^2 is just 0.0068, and none of these variables are statistically

significant predictors of growth. These results suggest that growth is likely driven by other factors, such as innovation, competitive positioning, or market dynamics. The correlation analysis of ownership and firm performance supports this conclusion, showing weak relationships between revenue growth and ownership or leadership variables. For instance, the correlation between revenue growth and CEO ownership is just 0.012, while the correlation with ROIC is -0.051. These findings indicate that growth is influenced by broader strategic and market-driven factors beyond internal ownership structures. Further insights into the impact of ownership on growth emerge from the analysis of CAPEX (Capital Expenditures) as a percentage of market capitalization. The results indicate that CEO ownership has a negative but marginally significant impact on CAPEX spending. This suggests that higher CEO ownership is associated with more conservative investment strategies, potentially reflecting a focus on operational efficiency rather than aggressive expansion. In contrast, executive ownership has no significant relationship with CAPEX.

Another Pearson correlation analysis reveals a weak negative correlation between CAPEX and CEO ownership (-0.074), further underscoring the cautious investment behavior of high-ownership CEOs. While this conservatism may enhance operational performance, it could also limit growth opportunities, particularly in competitive or fast-evolving markets. CEO ownership significantly enhances operational performance, as evidenced by its strong positive impact on ROIC and ROA (Tables A3, A4). However, it has limited influence on revenue growth and no meaningful effect on stock performance. High-ownership CEOs appear to adopt more conservative investment strategies, as reflected in lower CAPEX spending, which may enhance efficiency but constrain growth in dynamic markets. To maximize value, firms must carefully balance CEO incentives with strategic investment decisions, avoiding pitfalls such as entrenchment or underinvestment.

4.3 Impact of CEO Tenure and Age on Firm Performance

This section examines the relationship between CEO tenure and age on key firm performance metrics, specifically ROA, ROIC, and stock performance. The analysis rejects the null hypothesis (H_0: CEO age and tenure have no significant effect on firm performance) and supports the alternative hypothesis (H_1: CEO age and tenure are significantly associated with firm performance). Below is a detailed discussion of the findings, referencing all relevant tables. Each additional year of CEO tenure correlates with a 0.123% increase in ROA. This finding likely reflects how longer-tenured CEOs develop a deeper understanding of the firm's workings and can optimize resources over time. The positive correlation between tenure and ROA is further supported by a Pearson correlation coefficient of 0.1149 (Table A7). Similarly, CEO tenure contributes to a 0.192% improvement in ROIC for each additional year of tenure. The correlation between tenure and ROIC is weaker but still positive at 0.0849.

$$Stock\ Performance = 440.3 + 14.86\,(CEO\ Total\ Tenure) - 5.27 * (CEO\ Age)$$
$$ROA = 13.53 + 0.123 * (CEO\ Total\ Tenure) - 0.13 * (CEO\ Age)$$
$$ROIC = 21.42 + 0.192 * (CEO\ Total\ Tenure) - 0.204 * (CEO\ Age)$$

Despite these findings, the explanatory power of the model is limited. Tenure accounts for only 2.3% of the variance in ROA and 1.86% of the variance in ROIC (R^2 values from Table A5). However, as Cohen (1988) argues, small effect sizes can have substantial real-world implications, particularly in large samples like the ones analyzed here. For example, even minor improvements in ROA or ROIC can translate into significant financial impacts for large-cap corporations. CEO age shows a more complex relationship with performance. While older CEOs tend to exhibit more conservative decision-making, which may stifle innovation, their decisions can also stabilize operations. The correlation between CEO age and ROA is negative and similarly weak for ROIC. Descriptive statistics show that the average CEO age in the sample is 58.94 years, with a standard deviation of 6.56 years.

Table A5: Regression Coefficients – *Impact of CEO Tenure and Age on Firm Performance Metrics (ROA, ROIC, and Stock Performance)*

Model fit, significance, and observations for ROA, ROIC, and stock performance analysis.

Metric	R^2	Adjusted R^2	F-statistic	Significance F	Standard Error	Observations
ROA	0.023	0.0188	5.45	0.0046	7.9	465
ROIC	0.0186	0.0146	4.68	0.0097	13.69	496
Stock Performance	0.0133	0.009	3.1	0.0459	1205.57	465

Table A6: Regression Model Summary – *Impact of CEO Tenure and Age on Firm Performance Metrics (ROA, ROIC, and Stock Performance)*

Coefficients and significance of CEO tenure and age on ROA, ROIC, and stock performance.

Metric	Variable	Coefficient	Standard Error	t-Statistic	P-value	95% Confidence Interval
ROA	**Intercept**	13.53	3.39	3.99	0.0001	[6.87, 20.19]
ROA	**CEO Total Tenure at Company**	0.123	0.0395	3.11	0.002	[0.0454, 0.2007]
ROA	**CEO Age**	-0.13	0.0601	-2.16	0.0315	[-0.2477, -0.0116]
ROIC	**Intercept**	21.42	5.62	3.81	0.0002	[10.37, 32.47]
ROIC	**CEO Total Tenure at Company**	0.192	0.067	2.87	0.0043	[0.0604, 0.3239]
ROIC	**CEO Age**	-0.204	0.1	-2.04	0.042	[-0.4003, -0.0074]
Stock Performance	**Intercept**	440.3	517.19	0.85	0.395	[-576.03, 1456.63]
Stock Performance	**CEO Total Tenure at Company**	14.86	6.03	2.47	0.014	[3.02, 26.71]
Stock Performance	**CEO Age**	-5.27	9.17	-0.57	0.5657	[-23.28, 12.74]

The relationship between tenure and age introduces additional complexities. Older CEOs often have longer tenures, leading to potential multicollinearity. While this study does not explicitly test for multicollinearity, the positive correlation between age and tenure (0.3661, Table A7) suggests that their effects on firm performance are interdependent. For instance, a long-tenured, older CEO may contribute to stagnation, while a younger CEO with similar tenure may bring dynamism and adaptability. The analysis finds that longer CEO tenures correlate with better stock performance, likely due to the stability and strategic continuity they provide. However, CEO age has no statistically significant effect on stock performance. With an R^2 of just 1.33%, these findings suggest that broader market and industry dynamics play a much larger role in influencing stock returns than CEO demographics.

Table A7: Example of Multicollinearity – Correlation Analysis Between Long-Term ROA, CEO Tenure, and CEO Age

Correlation between long-term ROA, CEO tenure, and CEO age have minor interdependencies.

	LT ROA	*CEO Total Tenure at Company*	*Chief Executive Officer Age*
LT ROA	1		
CEO Total Tenure at Company	0.114879041	1	
Chief Executive Officer Age	-0.05026338	0.36610781	1

Real-world examples illustrate the interplay between CEO tenure, age, and firm performance. For instance, Adam Neumann of WeWork is an example of a young, inexperienced CEO whose erratic leadership and risky strategies led to significant destruction of firm value. This case highlights how markets demand higher risk premiums for young CEOs, which often diminish over time as they gain experience. Satya Nadella at Microsoft, a relatively young CEO with a transformative leadership style, has driven exceptional performance through innovation and strategic pivots. In contrast, Lee Raymond of ExxonMobil, an older CEO with a long tenure, maintained stability and operational efficiency but faced criticism for a lack of adaptability in dynamic market conditions. These examples support the idea that the impact of CEO age and tenure depends on their interaction with firm characteristics, industry dynamics, and market conditions.

Empirical evidence suggests that firm value often follows a hump-shaped pattern with CEO tenure. This phenomenon may stem from the interplay of age and tenure: while early-tenure CEOs drive growth through fresh perspectives and bold strategies, excessively long tenures can lead to stagnation as failures accumulate, or adaptability declines. This pattern is especially pronounced in dynamic industries where CEOs must adapt to rapid changes. Brochet et al. (2021) argue that the optimal tenure depends on factors such as industry volatility and labor market conditions, emphasizing that no one-size-fits-all approach exists. Older CEOs with longer

tenures tend to adopt more conservative investment strategies, as reflected in lower CAPEX spending. While this conservatism, reflected in lower CAPEX spending and consistent with Upper Echelons Theory, reflects tight budgets and risk aversion, it is often associated with longer tenures and older executives, which may improve efficiency but also limit growth opportunities, particularly in competitive markets.

This study has several limitations. For instance, the interplay between CEO age and tenure was not explicitly tested for multicollinearity. However, given that their correlation is $r = 0.3661$, well below the typical thresholds of 0.7 or 0.8 that indicate potential multicollinearity, this issue is unlikely to significantly affect the results (Cohen, 1988). Nonetheless, future research could formally address this interaction to further validate the findings. Future research could address this by using advanced statistical techniques, such as variance inflation factors (VIFs), to isolate the unique effects of each variable. Additionally, the study did not fully explore interaction effects between CEO characteristics and external factors like market dynamics or firm size. These interactions could explain the hump-shaped patterns observed in firm value. Incorporating variables such as CEO personality traits, strategic investments, or industry-specific conditions could provide a more nuanced understanding of the relationship between CEO demographics and performance. Future research could also explore additional case studies, such as a young CEO with a long tenure in a high-performing firm (e.g., Mark Zuckerberg at Meta) or an older CEO with a short tenure in a low-performing firm (e.g., Bob Chapek's brief and tumultuous tenure at Disney).

The analysis demonstrates that while CEO tenure consistently improves operational metrics like ROA and ROIC, the effects of age are more nuanced and often negative. Tenure and age together account for only a small percentage of the variance in firm performance metrics (e.g., 2.3% for ROA, 1.86% for ROIC, 1.33% for stock performance). Nonetheless, even small effects can have significant practical implications in large samples or firms with substantial market capitalizations. It is worth noting however that demographic indicators like age and tenure may introduce more noise than psychological measures, as they serve as indirect proxies for traits such as risk propensity, motivation, socioeconomic background and cognitive style. Future research should further explore the interplay of CEO characteristics with external and firm-specific factors to provide a more comprehensive view of their impact on firm performance.

5. Conclusions

This study analyzed the relationships between CEO compensation, ownership, and demographic characteristics, focusing on their influence on firm performance metrics such as ROIC, ROA, and stock performance. While measurable, the overall impact of these factors remains modest, as reflected in the low R^2 values. Although the overall impact is modest (reflected in low R^2 values), these findings remain meaningful, as they emphasize the complexity of firm performance and the importance of examining these variables more critically to derive actionable insights for corporate governance.

Regarding compensation, the findings confirm a correlation between high levels of pay and improved firm performance, especially when compensation growth is considered. However, the effect size is modest, underscoring that simply paying executives more is not a solution. Instead, the design of compensation packages emerges as a crucial element. Equity-based compensation aligns the interests of executives with those of the company, fostering long-term value creation by addressing the agent-principal dynamic. While there is evidence that higher pay can improve results, this relationship is not linear or straightforward. For practitioners, this means the focus should shift from the *amount* of pay to the *structure* of compensation. Boards should prioritize equity-heavy designs while avoiding incentives that encourage short-termism or excessive risk-taking. Future research could provide deeper insights into the optimal balance of equity-based compensation, identifying the optimal balance that maximizes executive effectiveness.

For ownership, the study finds that CEO ownership significantly improves operational metrics like ROIC and ROA but does not translate into better stock performance. This divergence may occur because markets discount high CEO ownership companies, or because high ownership leads to conservative decision-making that prioritizes cost control over innovation and growth. Ownership appears particularly effective in stable industries, where cost efficiency is a primary concern. However, in fast-evolving markets, this conservatism may hinder adaptability and R&D investments, limiting growth potential. For example, in dynamic sectors like technology, excessive ownership caution might stifle innovation and risk-taking, while in stable industries such as utilities, higher CEO ownership often fosters efficient cost management and long-term stability. Boards in dynamic industries should therefore balance CEO equity stakes with alternative incentive structures that encourage innovation and strategic agility.

Demographic characteristics, particularly tenure and age, also shape firm performance. Longer CEO tenures improve operational efficiency and stabilize operations through institutional knowledge and strategic continuity, as evidenced by their positive impact on ROIC and ROA. However, longer tenures and older age are also associated with greater conservatism, which can suppress innovation and risk-taking. While experienced CEOs optimize capital efficiency, their cautious approach may lead to stagnation. For firms in fast-changing industries, this conservatism could hinder competitive growth. Boards should not simply favor long-tenured or older CEOs but instead seek a balance that reflects the specific needs of the industry and the firm's vision. For innovative sectors, younger or shorter-tenured leaders may offer the fresh perspectives and risk tolerance necessary for success.

Across all three areas - compensation, ownership, and demographics - low R^2 values indicate that many factors beyond CEO characteristics significantly shape firm performance. This finding suggests that executive factors, while measurable, rarely account for the bulk of performance outcomes. Boards should avoid overreliance on any single variable, as firm success depends on a broader interplay of trends in the industry, market forces, and organizational quality. High-performing companies with strong fundamentals are more likely to succeed regardless of leadership design, while weaker companies may struggle even under exceptional

leaders. In conclusion, CEO compensation, ownership stakes, and demographic factors exert measurable but modest effects on firm performance. For boards and practitioners, the implications are clear: governance strategies must integrate these internal mechanisms with a broader understanding of external realities to sustain long-term growth and competitiveness. Future research should further investigate the contextual application of equity-based compensation and the interaction between CEO traits and external market conditions to better inform leadership effectiveness.

Works Cited

Aguilar, F. J. (1967). Scanning the Business Environment. New York: Macmillan.

Allen, L. (1979). Thoughts on the Business Environment. Harvard Business Review.

Amihud, Y., & Lev, B. (1981). Risk reduction as a managerial motive for conglomerate mergers. The Bell Journal of Economics.

Araujo-Cabrera, Y., Suarez-Acosta, M. A., & Aguiar-Quintana, T. (2017). Exploring the influence of CEO extraversion and openness to experience on firm performance: The mediating role of top management team behavioral integration. Journal of Leadership & Organizational Studies.

Bantel, K. A., & Jackson, S. E. (1989). Top management and innovations in banking: Does the composition of the top team make a difference? Strategic Management Journal.

Becker, G. S. (1970). Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education. New York: Columbia University Press.

Blonski, M., & Lilienfeld-Toal, U. (2010). Moral hazard with excess returns. Journal of Economic Theory.

Boyd. (1994). Board Control and CEO Compensation. Strategic Management Journal. 335-344

Brochet, F., Limbach, P., Schmid, M., & Scholz-Daneshgari, M. (2021). CEO tenure and firm value. The Accounting Review.

Carpenter, M. A., Geletkanycz, M. A., & Sanders, W. G. (2004). Upper echelons research revisited: Antecedents, elements, and consequences of top management team composition. Journal of Management.

Channon, D. F. (1979). The growth strategies of firms: Strategic, financial, and organizational correlates. Journal of Business Strategy.

Chatterjee, A., & Hambrick, D. C. (2007). It's all about me: Narcissistic CEOs and their effects on company strategy and performance. Administrative Science Quarterly.

Child, J. (1972). Organizational structure, environment, and performance: The role of strategic choice. Sociology.

Child, J. (1974). Managerial and Organizational Factors Associated with Company Performance. London Department of Employment.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates.

Collins, O. F., & Moore, D. G. (1970). The organization makers: A behavioral study of independent entrepreneurs. New York: Appleton-Century-Crofts.

Collins, R. (1971). Functional and Conflict Theories of Educational Stratification. American Sociological Review.

Core, J. E., & Guay, W. R. (1999). The Use of Equity Grants to Manage Optimal Equity Incentive Levels. Journal of Accounting and Economics.

Cremers, K. J., & Ferrell, A. (2014). Thirty years of shareholder rights and firm valuation. The Journal of Finance.

Crossland, C., & Hambrick, D. C. (2011). Differences in managerial discretion across countries: How nation-level institutions affect the degree to which CEOs matter. Strategic Management Journal.

Crystal, G. S. (1991). In search of excess: The overcompensation of American executives. W. W. Norton & Company.

De Hoogh, A. H., Greer, L. L., & Hartog, D. N. (2015). Diabolical dictators or capable commanders? An investigation of the differential effects of autocratic leadership on team performance. The Leadership Quarterly.

Devers, C. E., Wiseman, R. M., & Holmes, R. M. (2007). The effects of endowment and loss aversion in managerial stock option valuation. Academy of Management Journal.

Eisenhardt, K. M. (1989). Agency theory: An assessment and review. Academy of Management Review.

Fama, E. F., & Jensen, M. C. (1983). Separation of ownership and control. Journal of Law and Economics.

Frye, M. B. (2004). Equity-Based Compensation for Employees: Firm Performance and Determinants. The Journal of Financial Research.

Gabaix, X., & Landier, A. (2008). Why has CEO pay increased so much? The Quarterly Journal of Economics.

Geletkanycz, M. A., & Black, S. S. (2001). Bound by the past? Experience-based effects on commitment to the strategic status quo. Journal of Management.

Gibbons, R., & Murphy, K. J. (1990). Relative Performance Evaluation for Chief Executive Officers. Industrial and Labor Relations Review.

Gompers, P. A., Ishii, J. L., & Metrick, A. (2003). Corporate governance and equity prices. The Quarterly Journal of Economics.

Gorton, G., & Huang, L. (2014). Asset prices when agents are marked-to-market. National Bureau of Economic Research.

Guang-sheng, Y. (2010). Empirical investigation on the relationship between CEO tenure and firm performance in China. Commercial Research.

Hambrick, D. C. (2007). Upper echelons theory: An update. Academy of Management Review.

Hambrick, D. C. (2018). Upper Echelons Theory : Origins, twists and turns, and lessons learned. Emerald Handbook of Management and Organization Theory.

Hambrick, D. C., Cho, T. S., & Chen, M. J. (1996). The influence of top management team heterogeneity on firms' competitive moves. Administrative Science Quarterly.

Hambrick, D. C., & Mason, P. A. (1984). Upper Echelons: The Organization as a Reflection of Its Top Managers. The Academy of Management Review.

Harris, M. (1979). *Optimal incentive contracts with imperfect information.* Journal of Economic Theory.

Hart, P., & Mellons, J. (1970). *Management Youth and Company Growth: A Correlation?* Management Decision.

Hillman, A. J., & Dalziel, T. (2003). *Boards of directors and firm performance: Integrating agency and resource dependence perspectives.* Academy of Management Review.

Horton, J., Millo, Y., & Serafeim, G. (2012). *Resources or power? Implications of social networks on compensation and firm performance.* Journal of Business Finance & Accounting.

Jensen, M., Potočnik, K., & Chaudhry, S. (2020). *A mixed-methods study of CEO transformational leadership and firm performance.* European Management Journal.

Jensen, M. C., & Meckling, W. H. (1976). *Theory of the firm: Managerial behavior, agency costs and ownership structure*. Journal of Financial Economics.

Judge, T. A., & Piccolo, R. F. (2004). *Transformational and transactional leadership: A meta-analytic test of their relative validity.* Journal of Applied Psychology.

Kaufmann, D. (2004). *Governance redux: The empirical challenge.* In World Bank Institute Working Papers.

Kerr, J., & Bettis, R. A. (1987). *Boards of Directors, Top Management Compensation, and Shareholder Returns.* Academy of Management Journal.

Kurdyukov, N. (2023). *CEO Personal Traits and Company Performance: Evidence from Russia.* Journal of Corporate Finance Research.

Lawrence, P. R. (1997). *The black box of organizational behavior.* Academy of Management Perspectives.

Leland, H. E., & Pyle, D. H. (1977). *Informational asymmetries, financial structure, and financial intermediation*. Journal of Finance.

Liden, R. C., Wayne, S. J., Zhao, H., & Henderson, D. (2008). *Servant leadership: Development of a multidimensional measure and multi-level assessment.* The Leadership Quarterly.

Lin, J.-C., & Howe, J. S. (1990). *Insider trading in the OTC market*. Journal of Finance.

Malmendier, U., & Tate, G. (2005). *CEO overconfidence and corporate investment.* The Journal of Finance.

Markoczy, L. (1997). *Measuring beliefs: Accept no substitutes.* Academy of Management Journal.

McDonald, M. L. (2013). *Access denied: Low mentoring of women and minority first-time directors and its negative effects on appointments to additional boards.* Academy of Management Journal.

Mehran, H. (1995). *Executive compensation structure, ownership, and firm performance.* Journal of Financial Economics.

Miles, R. E., & Snow, C. C. (2003). *Organizational strategy, structure, and process.* Stanford University Press.

Morris, R. D., & Attaway, M. C. (2000). *The Effect of CEO Compensation on Firm Financial Performance*. Journal of Accounting, Auditing & Finance.

Neely, A., Gregory, M., & Platts, K. (2020). *Performance measurement system design: A literature review and research agenda.* International Journal of Operations & Production Management.

Nielsen, S. (2009). *Why do top management teams look the way they do? A multilevel exploration of the antecedents of TMT heterogeneity.* Strategic Organization.

Ormiston, M. E., Wong, E. M., & Ha, J. (2022). *The role of CEO emotional stability and team heterogeneity in shaping the top management team affective tone and firm performance relationship.* The Leadership Quarterly.

Priem, R. L., Lyon, D. W., & Dess, G. G. (1999). *Inherent limitations of demographic proxies in top management team heterogeneity research.* Journal of Management.

Report of Investigation of Enron Corporation and Related Entities Regarding Federal Tax and Compensation Issues, and Policy Recommendations. (2003). *Joint Committee on Taxation.* https://www.jct.gov/getattachment/d3202b28-9432-4270-ac48-7faf47b2997c/x-36-03-1764.pdf

Robertson, H. (2019, September 23). *WeWork's mess, explained.* Vox. https://www.vox.com/recode/2019/9/23/20879656/wework-mess-explained-ipo-softbank

Ruenzi, S., & Lilienfeld-Toal, U. (2014). *CEO Ownership, Stock Market Performance, and Managerial Discretion*. The Journal of Finance.

Schepker, D. J., Kim, Y., Patel, P. C., Thatcher, S. M., & Champion, M. C. (2017). *CEO succession, strategic change, and post-succession performance: A meta-analysis.* The Leadership Quarterly.

Sebbas, A. O. (2017). *CEO Cultural Background and its Relation to Companies' Performance in Europe.* Helda Helsinki.

Shleifer, A., & Vishny, R. W. (1997). *A survey of corporate governance.* Journal of Finance.

Tirole, J. (2001). *Corporate governance*. Econometrica.

Turner, J. C. (1982). *Towards a cognitive redefinition of the social group.* Cambridge University Press.

Urquhart, A. (2022). *PhD CEOs and firm performance*. European Financial Management.

Walters, B., Kroll, M., & Wright, P. (2007). *CEO tenure, boards of directors, and acquisition performance.* Journal of Business Research.

Wiersema, M. F., & Bantel, K. A. (1992). *Top management team demography and corporate strategic change.* Academy of Management Journal.

Taxing Carbon in the United States: A Socially Just, Political Economy Approach
By Isabell Luo

Abstract

Despite the global shift toward carbon pricing, the United States remains one of the few major economies without a federal carbon tax. To date, stakeholders from across the political economy have opposed carbon tax policies. This includes environmental justice organizations who seek to prevent inequitable application of such a tax as well as corporate actors who aim to protect their current operations and interests. This study seeks to bridge these gaps by proposing a cohesive carbon tax framework that balances efficient emissions reduction, protection for marginalized communities, and political feasibility. The paper introduces an EJ policy approach—WHO, HOW, WHAT—that identifies key stakeholders (WHO), methods for achieving equitable outcomes (HOW), and strategic solutions (WHAT). This framework is supported by original stakeholder mapping (see Fig. 1) to uncover the complex influences shaping U.S. carbon tax policy. Literature reviews on EJ, federal carbon tax design, and case studies from Ireland, Canada, California, Washington State, and the Northeast provide additional insights. The analysis establishes three key criteria for an effective, socially just, and politically viable carbon tax: clear communication, revenue recycling, and stakeholder engagement. This holistic approach offers actionable guidance for designing federal carbon tax policies.

1. Introduction

1A. Carbon Tax Literature Review

A carbon tax charges emitters per ton of CO2 and encourages businesses and individuals to cut emissions through efficient practices, cleaner fuels, or lifestyle changes to reduce tax costs (Partnership for Market Readiness 10). Economic theory suggests setting the carbon tax rate equal to the SCC, which would account for the total damage from an additional ton of CO2 (Hafstead "Carbon"). However, policymakers often set tax rates lower than the SCC due to uncertainty and the wide range of estimates, typically basing rates on politically feasible climate targets or revenue goals (Kaufman "What"). For any given tax rate, the carbon tax paid depends on the carbon content of fossil fuels, which correlates with CO2 emissions (Hafstead "Carbon"). For instance, with a $50 per ton CO2 tax, burning a ton of coal, which emits about 2.6 tons of CO2, would incur a $130 tax (Moseman and Surendranath). Setting the carbon tax too low undermines its effectiveness in reducing emissions, while setting it too high can lead to economic and political pushback. This section explores various aspects of carbon pricing, including its design, stakeholders, costs and benefits, real-world examples, and political feasibility

Carbon pricing relies on two inverse mechanisms: carbon taxes and cap-and-trade systems. A carbon tax charges a fee on the carbon content of fossil fuels, offering price stability for emitters but uncertain emissions reduction (Kaufman "Carbon"). Cap-and-trade caps total

emissions through limited allowances and lets the market set permit prices but cause price volatility ("Cap and Trade"). While carbon tax enforcement easily integrates with existing tax systems, cap-and-trade requires a costly regulatory framework for permit auctions, trading, and monitoring offset and credit compliance (Zakrzewski “Carbon Fee”; Aldy and Stavins 157-158). Cap-and-trade also risks greenwashing due to manipulation in permit trading and weak verification systems in linked markets (Diedrich). Both methods generate revenue, but carbon taxes yield more predictable and higher per capita revenue, supporting more stable funding for green initiatives and consumer investments (Carl and Fedor 53-54). Politically, cap-and-trade avoids the stigma of a direct tax but complicates communication with the policy’s complex permit trading system (Stavins 51). Though theory finds little difference between carbon taxes and cap-and-trade, practical challenges in regulation, greenwashing, revenue, and benefit communication make carbon taxes the more effective choice.

Carbon taxes efficiently address three major climate-related market failures (Ho). First, carbon taxes tackle the lack of accountability for environmental damage by internalizing CO2 costs, ensuring that prices reflect environmental impacts and drive efficient emissions reductions (Baranzini et al. 12). Second, carbon taxes address the under compensation of green companies by making carbon-emitting technologies more expensive. This rewards advancements in energy efficiency and innovation (Baranzini et al. 4). Third, carbon taxes shift behavior away from fossil fuels and use revenue to support a just transition for affected workers and EJ communities. (Baranzini et al. 3; Gazmararian 2). By gradually increasing tax rates, carbon taxes ensure long-term emissions reductions and bolster support for renewable energy.

Countries have adopted carbon pricing policies, but governments must take further action to meet climate targets. Major economies like the EU, Canada, Mexico, China, the UK, and Australia have implemented carbon pricing (World Bank Group). Currently, 75 carbon taxes and ETSs cover 24% of global emissions, but these prices fall short of Paris Agreement targets. Emissions are projected to be 36% above the 2°C limit and 55% above the 1.5°C limit by 2035 (World Bank 12). Carbon tax rates range from $0.46 to $167 per ton, with only 6 countries exceeding the $75 per ton needed for a 2°C limit, and none meeting the IPCC’s recommended $170 per ton by 2030 for a 1.5°C limit (World Bank Group; Black et al.; “Price Greenhouse”). Despite a record $104 billion in revenue from carbon pricing in 2023, $1.3 trillion in fossil fuel subsidies overshadows this (World Bank 28, 15). Enhancing carbon pricing requires reducing fossil fuel subsidies and increasing clean energy support, demanding strong political will and international coordination.

In recent years, U.S. federal carbon pricing efforts have repeatedly faltered despite some state-level successes. For instance, the Waxman-Markey bill, which aimed to establish a nationwide cap-and-trade system, passed the House in 2009 but never reached a Senate vote due to a threatened filibuster (Weiss). In 2018, the House introduced the Energy Innovation Carbon Dividend Act (EICDA), proposing a $15 carbon tax with revenue redistributed to citizens ("Energy Innovation"; United States, Congress, House, Energy Innovation 9, 32). While the EICDA failed to pass, the bill has been reintroduced with bipartisan backing, securing 95 House

cosponsors in the 117th Congress and becoming the most popular carbon pricing bill in American history ("Energy Innovation"; "Cosponsors: H.R.2307"; Ye “Carbon Pricing”). Currently, the only other carbon pricing proposal in the 118th Congress is the Market Choice Act, which starts at $35 and invests revenue in clean energy infrastructure and climate adaptation ("Carbon Pricing Bills"). Amid political polarization, the fate of these proposals remains uncertain, underscoring the need to examine the social, political, and economic barriers to carbon taxation.

1B. Environmental Justice Framework

Although there is a general consensus among scientists and economists that carbon pricing is the most economically effective method to reduce greenhouse gas emissions, such policies have faced criticism from environmental justice (EJ) groups. While a carbon tax could lower emissions and improve air quality, historical environmental marginalization has eroded trust in political and economic systems. This distrust raises concerns that offsets and carbon credit markets may increase land prices, displace communities, and threaten Indigenous land sovereignty ("Carbon Pricing Is"). Additionally, carbon taxes can disproportionately impact low-income households by increasing energy and goods costs without adequately reducing local pollution (Patterson et al. 98).

Despite these EJ criticisms, the core issue lies not in the concept of pricing carbon but in how such policies are designed—specifically, how stringent they are and how the revenue is allocated. This opens the possibility of designing an environmentally effective and politically feasible carbon tax that bridges the concerns of environmental justice advocates with the support of economists and the mainstream environmental movement.

This can be achieved by adopting an EJ policy framework to guide carbon tax design. An effective EJ policy framework integrates principles from EJ and climate justice (CJ) organizations into systems analysis through three key components: WHO, HOW, and WHAT.

- WHO: Identifies key stakeholders and agents through actor mapping. Inclusive participation is essential for legitimacy and trust, bringing diverse voices to the table.
- HOW: Determines real-world methods for achieving goals, incorporating conflict and resilience analysis to ensure a legitimate process and sustainable solutions.
- WHAT: Develops strategic approaches to problem-solving, focusing on actionable results while maintaining trust. The stakeholders identified in the WHO stage must drive the WHAT solutions to ensure coherence and effectiveness.

Applying these principles to climate policy, EJ groups emphasize reducing pollution, improving health outcomes, providing financial support for low-income households, and involving EJ communities in policy design and implementation. Integrating EJ and CJ principles into carbon pricing design can help build stronger coalitions within the broader environmental movement, fostering inclusive decision-making and equitable outcomes. By using this

comprehensive approach, policymakers can create carbon tax policies that balance environmental protection with social equity, gaining broader support. The rest of this paper will further develop this framework to analyze stakeholder considerations in carbon tax policy, draw lessons from five case studies on existing carbon pricing mechanisms, and examine the criteria for a federal carbon tax.

2. Stakeholder Mapping

The political economy of a carbon tax hinges on how costs and benefits distribute, influencing public opinion and political support. This section outlines the key players involved in carbon tax policy, examining their interests, influence, and potential conflicts. Categorizing stakeholders into corporate entities, policymakers, and key constituencies reveals the diverse perspectives and power dynamics shaping the debate. This analysis provides a foundation for navigating the complexities of implementing an effective and equitable carbon tax. Figure 1 provides a visual summary of the stakeholder mapping in the following section.

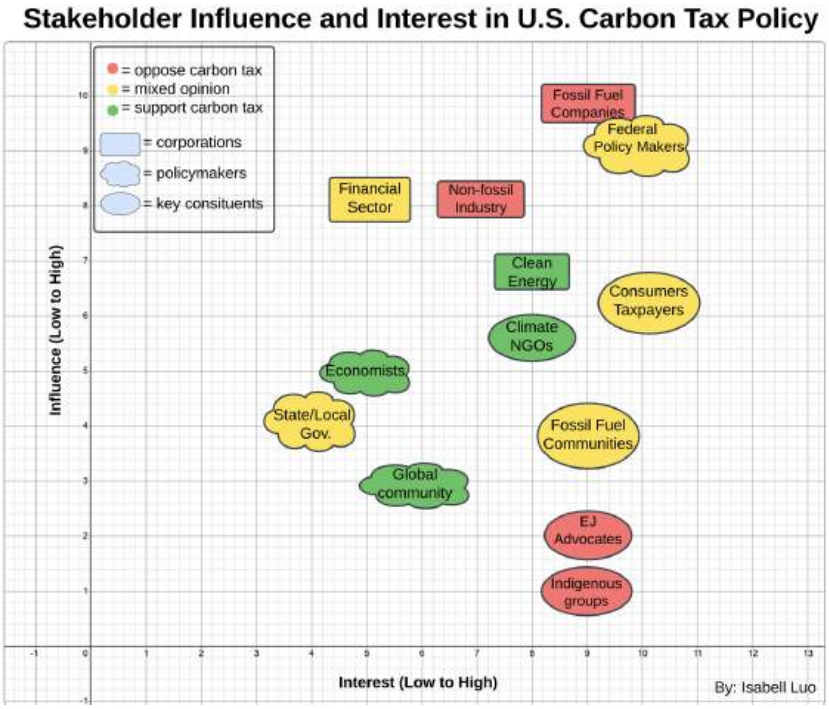


Figure 1. Stakeholder Influence and Interest in U.S. Carbon Tax Policy

Note: This map is not to scale and was modeled to reflect existing carbon pricing proposals and implemented mechanisms in the United States. Stakeholder positions are not fixed and may shift depending on the specific design of each policy.

2A. Corporations

1. Fossil Fuel Companies

- Interests: Fossil fuel companies traditionally opposed carbon taxes due to their impact on profits. Yet, major firms like ExxonMobil, Chevron, BP, and Shell now back carbon pricing to weaken coal competition, exploit inelastic demand, anticipate carbon leakage, and enhance their climate image (Naef 1). This support often acts as greenwashing, allowing them to avoid significant financial losses while opposing stringent climate policies and influencing lenient pricing mechanisms (Pulkkinen). They also spread misinformation and engage in regulatory capture, such as undermining the SEC's climate disclosure rules (Martinez et al. 4).
- Power dynamics: Through extensive lobbying by groups like the American Petroleum Institute (API), fossil fuel companies secure tax breaks and policies that increase emissions (United States, Congress, House, Committee on Oversight and Reform 8). This lobbying weakens the effectiveness of carbon taxes, especially when taxes are poorly enforced and counteracted by subsidies.
- Potential Conflicts: These companies often conflict with climate advocates, EJ groups, and policymakers seeking stronger carbon pricing and subsidy reforms.

2. Industrial Sectors (Excluding fossil fuel companies)
    - Interests: Energy-intensive industries, such as chemicals, cement, and steel, fear competitiveness issues and carbon leakage from carbon pricing ("Carbon Tax Basics"). Other companies like Microsoft use internal carbon pricing to manage risks and save costs, with sector-specific rates (Chris Fawsona et al. 17). Carbon taxes often exempt agricultural fuels and non-fossil emissions from livestock, easing agricultural opposition (Dumortier and Elobeid 1).
    - Power Dynamics: These industries exert influence through trade associations, lobbying, and media.
    - Potential Conflicts: Disputes may arise between industry groups and environmental advocates over stricter carbon pricing, and differing internal pricing strategies may affect views on federal policies.
3. Clean Energy Sector and Green Innovation
    - Interests: The clean energy sector supports carbon pricing to level the playing field with fossil fuels. Carbon taxes could boost renewables to 29-41% of U.S. electricity by 2030 and reduce coal production by 28-84% (Larson et al. 7).
    - Power Dynamics: Clean energy has grown profitability, with $97 billion invested in the U.S. in 2022 ("World Energy"). Despite increased economic influence, the industry still faces a green premium depending on geography and energy sector.
    - Potential Conflicts: Conflicts may occur over transition speed, revenue use, and market competition with fossil fuels. Initially, carbon taxes may reduce renewable energy use due to labor and technology adaptation costs, but after about three years eventually promote innovation and growth (Yin et al. 2, 7).
4. Financial Sector Players: (Including banks, investment firms, etc.)

- Interests: With increased ESG and carbon pricing initiatives, the financial sector focuses on long-term risk management and profitability in carbon credit markets, clean energy investments, and green bonds ("OJK International").
- Power Dynamics: The financial sector influences market trends and policies through investment decisions and intermediary roles.
- Potential Conflicts: Investors increasingly push for sustainability, as seen with board changes at ExxonMobil and Chevron's vote to cut Scope 3 emissions in 2021 (Hiller and Herbst-Bayliss; Reuters). Despite growth in sustainable finance, greenwashing and fossil fuel investments persist ("The Rise"). Banks often reduce domestic lending to fossil fuels under carbon taxes but increase foreign lending in countries with weaker regulations, weakening environmental benefits (Laeven and Popov 1).

2B. Policymakers and policy advisors

5. Federal Policymakers
    - Interests: Policymakers aim to balance environmental goals with economic growth, social equity, and global competitiveness while managing subsidies to emissions-heavy sectors (Ye *Options* 3-4).
    - Power Dynamics: Policymakers shape the carbon tax's structure and enforcement, influenced by public opinion and their own reelection needs (Partnership for Market Readiness 41)
    - Policymakers face conflicts among industry lobbyists, environmental advocates, and constituencies. Despite Biden's efforts to cut fossil fuel subsidies, these financial supports persist due to over $30 million in lobbying since 2020, saving the industry $1.7 billion in 2025 alone and up to $15.6 billion over the next decade (Friedman). This regulatory capture of the U.S. Chamber of Commerce also complicates the successful implementation of carbon taxes and clean energy policies (Friedman).
6. State and Local Governments
    - Interests: Support for a federal carbon tax varies by region. California advances carbon pricing, but partisan divides shape federal carbon tax support with 15 Republican senators opposing such policy, citing potential harm to manufacturing and economic growth (Bill Cassidy, M.D.; Farber; United States).
    - Power Dynamics: States shape federal policy implementation based on local interests (Farber).
    - Potential Conflicts: States may conflict with federal policies if their climate rules are stricter or if less supportive states oppose federal mandates, leading to legal disputes and polarization.
7. Research and Academia (Economists)

- Interests: Economists laud carbon pricing as the most effective method to address environmental costs and promote climate action (Sterner et al. 1).
- Power Dynamics: Environmental economics has expanded but remains outside the mainstream (Sterner et al. 2). Economists shape policy through research but face challenges from poor communication, lobbying, and political resistance.
- Potential Conflicts: Disagreements exist within the economic community on specific carbon pricing approaches (Sterner et al. 6). Economists face challenges from industry lobbying and public skepticism regarding the fairness and effectiveness of carbon taxes.

8. International Organizations and Other Countries
    - Interests: International groups like the WTO, World Bank, OECD, and IMF advocate for carbon pricing and global taxing practices through international agreements such as the Paris Agreement (Partnership for Market Readiness 9-10; "WTO Chief"). These entities focus on regulating carbon credits and EU's Carbon Border Adjustment Mechanism (CBAM), which affect other countries by impacting global trade (Bellora and Fontagné 1; Smith).
    - Power Dynamics: They set global norms but have limited direct influence on U.S. climate policy.
    - Potential Conflicts: Their recommendations may clash with U.S. policies or affect national competitiveness.

2C. Key Constituents

9. EJ Advocates: including Climate Justice Alliance, 350.org, Sunrise, etc.
    - Interests: EJ advocates generally oppose carbon pricing due to spiritual and equity concerns, emphasizing the need to protect marginalized communities and recycle revenue for resilience (Boyce et al. 1).
    - Power Dynamics: Historically marginalized, EJ gains influence through grassroots efforts and national initiatives like Justice40 ("Justice40").
    - Potential Conflicts: They clash with mainstream environmentalists, policymakers, and fossil fuel companies when social equity issues are overlooked.
10. Indigenous Communities
    - Interests: They push for policies that respect tribal land sovereignty and support Indigenous-led climate initiatives ("Carbon Pricing Is").
    - Power Dynamics: Historically marginalized, they are gaining political power through legal rights and grassroots efforts with groups like the Indigenous Environmental Network.
    - Potential Conflicts: They may conflict with fossil fuel companies, policymakers, and climate advocates if policies fail to address Indigenous rights or worsen inequalities. Fossil fuel companies use lawsuits to silence Indigenous activists and

support organizations that engage in voter suppression, blocking climate policy progress (Martinez et al. 5).

11. Climate Advocates (Including Environmental NGOs)
    - Interests: Major environmental NGOs, like Citizens' Climate Lobby (CCL), back carbon taxes to reduce emissions and advocate bipartisan solutions ("Who Supports").
    - Power Dynamics: These NGOs influence public opinion and policy through grassroots lobbying but face challenges due to climate polarization and limited capacity compared to the fossil fuel lobby (McFarlane).
    - Potential Conflicts: They may conflict with fossil fuel industries and sometimes EJ advocates. Groups like the Climate Leadership Council (CLC), funded by Big Oil, push for less stringent carbon pricing, leading to concerns about greenwashing ("Partners"; "The Climate"; Lowenstein).
12. Average American Taxpayers and Consumers
    - Interests: They worry about how carbon pricing impacts energy costs and access to clean energy alternatives (Fremstad et al. 1; Marshall 14).
    - Power Dynamics: They wield influence through voting, public opinion, and purchasing power, though fossil fuel trade associations significantly outspend environmental NGOs to gain voter support, with a funding ratio of approximately 27 to 1 (Fremstad et al. 2; Downie and Brulle).
    - Potential Conflicts: They may struggle between supporting environmental policies and managing the financial impact of fossil fuel taxes. Revenue recycling could shift support (Fremstad et al. 6).
13. Fossil Fuel Communities (Workers and residents dependent on fossil fuel industry)
    - Interests: Fossil fuel communities seek job security and economic support, favoring initiatives that fund pensions, healthcare, and retraining (Gazmararian 2). Despite perceptions of resistance to government intervention, 66% would support climate policies with just transition assistance (Gazmararian 2). Education about declining coal competitiveness can also shift support toward clean energy (Gazmararian 1).
    - Power Dynamics: These communities, often in conservative states, can influence bipartisan support for a carbon tax by lobbying legislators. Overcoming misinformation and polarization, intensified by fossil fuel corporations, is essential for effective policy engagement.
    - Potential Conflicts: Conflicts may arise with environmental advocates and policymakers pushing for high carbon tax without just transition for fossil fuel communities.

The stakeholder mapping for carbon tax policy reveals a complex interplay of interests across three broad categories: corporate entities, policymakers, and key constituencies. Corporate entities, including fossil fuel companies, industrial sectors, the financial sector, and the clean

energy industry, have varied stakes in carbon pricing, ranging from opposition due to revenue impacts to support for its long-term benefits. Policymakers at federal, state, and local levels, alongside research and academia, hold significant power in shaping and implementing carbon tax policies, often balancing economic growth, social equity, and environmental goals. Key constituencies, including EJ advocates, Indigenous communities, climate NGOs, average American consumers, and fossil fuel communities, bring critical perspectives on equity, economic impacts, and support for a just transition. Balancing these diverse, and often conflicting interests is essential for designing and implementing a carbon tax that achieves emissions reductions, political viability, and equitable outcomes.

3. Five Carbon Pricing Case Studies

While the United States has not adopted a national carbon tax, examining the experiences of Ireland and Canada, along with state-level cap-and-trade programs in California, the Northeast/Mid-Atlantic (RGGI), and Washington State, offers valuable insights. These five case studies analyze the development, climate impact, economic effects, revenue recycling, distributional outcomes, political economy, and EJ aspects of these initiatives, providing key takeaways for a potential federal carbon tax.

3A. Ireland

Ireland serves as a relevant case study for carbon tax implementation in the United States due to its comparable political economy, marked by high economic development, inequality, and fossil fuel dependence. For example, both Ireland and the United States rank in the top 10 globally for GDP per capita but also have high economic inequality (above 0.5 on a scale of 1), exceeding most other developed nations ("GDP per Capita"; "Gini Coefficient"). Additionally, both countries rely heavily on fossil fuels, comprising about 83% of Ireland's and 84% of the United States' energy consumption in 2023 ("National Energy"; "U.S. Energy"). Despite these economic conditions and Ireland's heavy reliance on fossil fuels for home heating and high-emission transportation, the country successfully launched a carbon tax in 2009, starting at 15 euros/tCO2 (Muth 7; Parliamentary Budget Office 6). Ireland gradually broadened the tax to include more fossil fuels and raised the rate annually, reaching 56 euros/tCO2 in 2024—one of the top 10 highest carbon prices globally—with a goal of 100 euros/tCO2 by 2030 (Parliamentary Budget Office 2; World Bank Group). Ireland sustained a relatively high carbon tax over the long term by focusing on mitigating distributional impacts both horizontally (geographical and situational) and vertically (income-related) (Muth 9). In fact, about 30% of the carbon tax revenue supported additional public assistance for low-income households and vulnerable groups, such as large families and the elderly (Muth 9); the remaining 70% funded energy efficiency programs and the development of a low-carbon transport system (Muth 9). This hybrid approach, combining direct household benefits with climate-focused spending, secured political support for a carbon tax by aligning the interests of both public and private stakeholders (Muth 9). Thus, Ireland's success with hybrid revenue recycling serves as a

valuable model for addressing political economy and distributional impact challenges in U.S. carbon tax policy.

Key differences remain between Ireland and the U.S. Unlike the standalone national policy proposed in the U.S., Ireland's carbon tax works alongside the EU cap-and-trade system by taxing fossil fuels not covered by the EU policy and imposing an additional tax on some natural gas and coal (Parliamentary Budget Office 7). With the EU system in place since 2005, Ireland's public and corporations have had nearly 20 years to adapt to a well-established carbon pricing system ("Development of EU ETS (2005-2020)"). In contrast, carbon pricing proposals in the United States have repeatedly failed, and existing climate policies face backlash (Doniger; Wolman et al.). Therefore, analyzing a carbon pricing system in a country more comparable to the U.S. in terms of climate policy landscape is essential for developing transferable political messaging strategies.

3B. Canada

Canada provides a valuable case study for U.S. carbon tax policy due to similar government structures, slow federal environmental action, and regional capabilities for costly climate policies (Karapin 146). British Columbia's carbon tax, implemented in 2008 and leading to a federal tax in 2019, offers public support insights for potential U.S. policies. Both countries have comparable per capita emissions—14 to 15 tons of CO2—and are major global polluters, with Canada emitting 582 million tons and the U.S. 4.85 billion tons in 2022 ("CO2 Emissions"). Although not the largest energy producer, British Columbia produced 35% of Canada's natural gas in 2020 ("Provincial and Territorial"). This study will explore how BC maintained its carbon tax and the challenges facing Canada's federal tax.

British Columbia's revenue-neutral carbon tax in 2008, began at C$10 per ton and increased to C$30 by 2012, with tax cuts and industry to build support (Karapin 147-148; Fairbrother and Rhodes 2). However, public opposition grew before the 2009 elections due to concerns about rural areas, industry leniency, and gas prices though support rebounded with economic recovery from the 2008 recession (Karapin 148; Fairbrother and Rhodes 5). By 2015, support had reached 61% due to economic recovery and a 2.4% reduction in per capita emissions by 2011 (Karapin 149; Fairbrother and Rhodes 5). Since 2019, all Canadian jurisdictions have adopted carbon pricing, either through the federal tax or their own systems to meet national emissions standards ("Carbon Pollution"). Tax proceeds stay within the jurisdiction to address local needs, supporting households, businesses, and energy efficiency programs that aim to build broad support ("Canada Federal"; "Carbon Pollution").

Yet, Canada's federal carbon tax currently faces significant political challenges again due to inadequate public communication. Only half of Canadians recognize the Climate Action Incentive Payment (CAIP) rebate, and 44% hold negative views of the tax (Coletto). Financial concerns are significant, with 63% feeling rebates lag behind increased costs (Coletto). These negative perceptions increase in fossil fuel-dependent regions like Ontario and Alberta ("Additional Statistics"; Coletto). The Conservative Party has exploited this dissatisfaction to

oppose the tax, leading to decreased support for Trudeau, including a withdrawal of support from one in four Canadians and 10% of former Liberal voters. In response, Trudeau plans to rename the CAIP to Canada Carbon Rebate and improve direct deposit labeling that better communicates the rebate's link to carbon pricing, although he continues to face challenges in a fragmented media environment (Bonasia; Wherry).

The United States can learn from Canada's carbon pricing experience when considering a federal carbon tax. Canada's political climate highlights the urgent need for clearer communication about the carbon tax and rebate benefits to address misconceptions, reduce opposition, and secure long-term support for the policy. However, Canada's population, energy mix, and regional conditions differ from those in the U.S., limiting direct comparisons. Therefore, the next three case studies examine several subnational carbon pricing initiatives in the United States for more relevant context.

3C. New England and Mid-Atlantic, U.S.

Established in 2005, the Regional Greenhouse Gas Initiative (RGGI) is a coalition of eleven New England and Mid-Atlantic states aiming to reduce CO2 emissions from power plants through a cap-and-trade system ("RGGI 101 Factsheet"). Power plants over 25 MW must hold allowances for their emissions, with offsets covering up to 3.3% of obligations ("RGGI 101 Factsheet"). Each state manages its CO2 Budget Trading Program, setting limits, issuing allowances, and participating in auctions ("Elements of RGGI"). RGGI stabilizes the market with a price floor and ceiling, directs revenues to energy efficiency and renewable energy projects, and conducts regular reviews to ensure policy effectiveness and public input ("Elements of RGGI"; "RGGI 101 Factsheet").

RGGI's model delivered emissions reductions, health benefits, and economic growth to participating states. Since its inception, RGGI has cut power sector emissions by 50%, almost double the national rate, saving 300 to 830 lives, preventing over 8,200 asthma attacks, and avoiding 39,000 lost workdays, with $5.7 billion in health benefits (Manion et al.; "RGGI 101 Factsheet"). The program has also protected economic growth, with RGGI states' GDP rising by 47% since 2008, exceeding the national average by 31%, and generating $5.7 billion in economic benefits and 48,000 job-years (The Regional 2; Stuart and Hibbard 10). RGGI auctions have raised over $7 billion for investments in efficiency, clean energy, and consumer assistance, which reduced electricity costs by 5.7% compared to an 8.6% increase elsewhere ("RGGI 101 Factsheet"; The Regional 8). In 2022, RGGI states invested $364 million of auction revenue, projected to cut 7.5 million tons of CO2 and save $1.8 billion in energy costs for 246,000 households and over 2,600 businesses ("The Investment").

Despite RGGI's overall progress, net emissions reductions have not equally benefited EJ communities. While the shift from coal to natural gas has decreased overall emissions, a study from 1990 to 2015 finds that people of color and poorer groups are 15-25% more likely to live near a power plant (Declet-Barreto and Rosenberg 7). These EJ communities also carry more natural gas plants, leading to higher exposure to pollutants like sulfur dioxide and nitrogen oxide

(Declet-Barreto and Rosenberg 2). These health disparities highlight the need to ensure that future climate policies do not perpetuate inequities in energy project siting. To address these inequities, RGGI states are now incorporating EJ and equity into their policy review ("Program Review"). Recommendations include monitoring air quality, involving EJ communities in policy input, and directing funds to climate resilience projects in overburdened areas (Stuart and Hibbard 11-13). These steps aim to ensure a fairer distribution of environmental and health benefits.

RGGI's success in public engagement has led the Natural Resources Defense Council to call it a "model for the nation" due to its effectiveness and broad bipartisan support ("The Regional"). A 2016 survey found that 76% of residents in nine RGGI states view climate change as a serious issue, and 77% support their state's participation, citing benefits like improved air quality, public health, and lower electricity costs (Hart Research Associates and Chesapeake Beach Consulting 1-2). Furthermore, 79% of voters supported tightening the carbon cap to cut emissions by 5% annually, with 72% maintaining support even after counterarguments, showing strong bipartisan backing for more ambitious climate action (Hart Research Associates and Chesapeake Beach Consulting 2, 4).

Despite strong public support, maintaining participation in RGGI faces political hurdles. In 2022, Pennsylvania joined RGGI with 72% public approval but faced legal challenges from power producers and unions, who argued it was an unconstitutional tax (Rosen; Warner). The Commonwealth Court voided the rule in 2023, but environmental groups appealed to the Pennsylvania Supreme Court, which in 2024 allowed their intervention to defend RGGI under the Environmental Rights Amendment (Rosen; Martinez). Meanwhile, Virginia, the first Southern state to join RGGI in 2020, saw $800 million in revenue and a 22% reduction in emissions by 2023 ("Virginia Fails"; Hines-Acosta). However, in May 2024, Governor Youngkin's budget excluded RGGI, cutting funding for flood mitigation and energy efficiency, and a lawsuit challenged this exit (Hines-Acosta). These developments underscore the difficulties in implementing and sustaining carbon pricing amidst legal and political conflicts.

The RGGI case study offers key insights for a federal carbon tax. Its effective interstate collaboration, public engagement, and limited use of offsets (capped at 3.3%) ensure environmental accountability and progress. Reinvesting revenue in energy efficiency, clean energy, and environmental justice initiatives supports job creation, reduces energy costs, and builds public support. Political and legal challenges, like those in Pennsylvania and Virginia, show the difficulties of maintaining such programs against fossil fuel industry opposition and suggest the need for corporate concessions. However, RGGI's focus on the power sector does not address the diverse economic and energy needs across the U.S., where a federal tax would need to include sectors like transportation, agriculture, and industry. Nonetheless, RGGI's success in allowing states to allocate revenue based on local priorities suggests that a federal carbon tax could offer states similar autonomy, improving political feasibility and effectiveness.

3D. California, U.S.

In 2013, California introduced a cap-and-trade program aiming for an 80% reduction in greenhouse gas emissions by 2050 ("California Cap and Trade"). The carbon price floor started at $10 per ton, rising to $15 by 2017. This increased gasoline prices by 11–15 cents, but $50-$70 annual household rebates largely offset these costs (Karapin 156). The program, covering 85% of emissions, funds the Greenhouse Gas Reduction Fund with 35% for disadvantaged communities ("California Cap and Trade"). To prevent economic carbon leakage, the program provides free allowances for half the emissions from oil and gas industries, electricity and fuel must buy all allowances at auction (Karapin 155; "ARB Emissions"). Offsets are capped at 8%, with strict reporting and verification, and the policy links with Quebec's carbon market system to enable cost-effective trading ("California Cap and Trade").

California's cap-and-trade program has also spurred environmental and economic progress, investing $28 billion in climate projects, creating 30,000 jobs, and reducing emissions 5.3%–equivalent to removing 80% of the state's gas-powered cars (Governor Gavin Newsom). Moreover, the revenue has supported 420,000 zero-emission vehicle rebates, 22 million trees planted, 12,606 affordable housing units, and over 1,300 public transit projects (Governor Gavin Newsom). In 2023, new projects aimed to cut an additional 14.7 million metric tons of greenhouse gases (Governor Gavin Newsom).

Cap-and-trade has yielded mixed emissions results in California. Even though 76% of auction funds benefited low-income communities, research has increased scrutiny on local air quality impacts of cap-and-trade. A 2022 USC paper argues that GHG and co-pollutant emissions decreased more in wealthier areas, with less improvement in EJ communities (Pastor et al. 33).The California Air Resources Board (CARB) claims that no evidence points to cap-and-trade worsening air quality in EJ communities, attributing issues to broader factors like post-2008 economic recovery ("FAQ Cap-and-Trade").CARB addressed pollution inequalities by enforcing diesel truck regulations (2009-2015) that cut NOx by 70%, black carbon by 73%, and particulates by 74% in areas like West Oakland ("FAQ Cap-and-Trade"). The cap-and-trade program reduced electricity sector emissions by 48% as natural gas use fell from 61% in 2012 to 43% in 2019 (Lessmann and Kramer 4). However, the industrial sector saw a 6% emissions increase by 2019 due to free allowances reducing incentives (Lessmann and Kramer 5, 7). To mitigate these disparities, research suggests facility-specific caps and "no trade" zones for EJ communities (Roy and Burtraw; Pastor et al. 34).

The California cap-and-trade case study provides valuable insights for a federal carbon tax. Effective carbon pricing can significantly cut emissions, particularly in the electricity sector, while boosting economic growth and supporting low-income communities through investments in clean technologies and infrastructure. However, the program's limited impact on the industrial sector and EJ air quality highlights the need for a more targeted policy approach.

3E. Washington State, U.S.

Washington's cap-and-invest program, launched in 2023, targets a 95% reduction in emissions by 2050, covering 70% of the state's emissions across energy, industry, buildings, and

transportation ("Washington's Cap-and-Invest"). Despite allocating 35% of cap-and-trade revenue to EJ initiatives, Washington's program faces criticism. In 2021, 39 climate justice groups argued that cap-and-trade is "inferior to straightforward taxes" due to inconsistent revenue and limited reductions (Climate Justice Advocates 1-2). They criticized major oil companies for profiting from carbon credits and noted California's cap-and-trade has achieved only 5% of its 2030 target (Climate Justice Advocates 1). The groups claim the program increases energy costs for low-income drivers while allowing polluters to use offsets instead of cutting emissions directly (Climate Justice Advocates 1; Mesa). Concerns include potential market linkages with California and Quebec that could undermine effectiveness and fail to benefit overburdened communities (Aronoff). The Environmental Justice Council suggests delaying linkage until clear benefits emerge, warning that a larger market might weaken incentives and encourage offset hoarding (Mesa). Despite objections, the Washington Department of Ecology focuses on long-term stability and investment in frontline communities (Mesa). Advocates call for a more equitable, community-driven approach (Climate Justice Advocates 2).

EJ opposition to the cap-and-trade program follows several failed carbon tax proposals. Initiative 1631, a 2018 effort to impose a $15 per metric ton carbon fee increasing over time, failed despite support from environmental, labor, and Indigenous groups (Schimel). Washington's history of failed carbon tax proposals since 2013 continued despite concessions by Governor Inslee to the oil and gas industries (Schimel). BP's refusal to support the bill under the lenient Climate Leadership Council plan and its $13 million campaign against Initiative 1631, part of a $31.2 million oil and gas industry effort, contributed to its defeat by 57% (Bernton; Schimel; Lavelle). This underscores the difficulty of enacting climate policies when oil companies oppose concrete action despite supporting carbon pricing in theory. This underscores the difficulty of enacting climate policies when oil companies support carbon pricing in theory but oppose concrete legislative action.

Despite past failures to pass a carbon tax, Washington's 2021 cap-and-trade program gained unexpected support from BP, which had historically opposed carbon pricing (Yoder). The program, requiring companies to buy pollution permits and invest in carbon sequestration, aligns with BP's interests through its stake in Finite Carbon (Aronoff). BP's shift appears driven by profit motives, exploiting carbon offsets that may allow continued pollution without real environmental gains (Aronoff).

3F. Case Study Insights

The case studies from Ireland, Canada, RGGI, California, and Washington reveal key aspects of carbon pricing strategies. Ireland's carbon tax serves as a strong U.S. model, focusing on revenue neutrality and hybrid recycling to balance economic and social impacts. Canada's federal tax encountered backlash due to poor communication about rebate benefits. RGGI demonstrates the benefits of a multi-state framework for emissions reductions and economic growth but remains limited by its focus on the power sector. California's cap-and-trade program

showcases the potential of broad-based pricing but struggles with uneven effectiveness and EJ concerns. Washington's cap-and-invest system faces challenges from carbon offsets and unexpected support from Big Oil, leading to criticisms and repeal efforts. Lessons from these cases suggest four key strategies for a U.S. federal carbon tax:

1. Interstate Collaboration and Flexibility: Inspired by RGGI, a federal carbon tax should encourage state collaboration and flexibility, enabling revenue reinvestment that benefits local communities, drives economic growth, and promotes competitiveness.
2. Coverage and Regulatory Framework: Drawing from California, a federal carbon tax must cover all key sectors—power, transportation, industry, etc.—while enforcing a robust regulatory framework. Tailor policies to sectors and manage policy mix to prevent the waterbed effect, ensuring broad emissions reductions.
3. Equity and Public Engagement: Insights from RGGI, California, and Washington highlight the need to address EJ concerns with targeted actions, like facility-specific caps and incentives for cuts in overburdened communities, ensuring fair pollution reductions and wider public support.
4. Communication and Revenue Recycling: Lessons from Ireland and Canada stress the need for clear communication about carbon pricing benefits and strategic revenue recycling to offset costs for overburdened households and businesses. A federal carbon tax should adopt these strategies to enhance transparency, public benefits, and political viability.

4. Hypothetical Federal Carbon Tax Literature Review

Building on lessons from the case studies, the following section reviews literature on emissions reductions, competitiveness, economic impact, distributional effects, and political economy remedies for a hypothetical U.S. federal carbon tax.

4A. Emissions, Health, Carbon Offsets

A carbon tax reduces overall emissions by charging per ton of CO2, affecting sectors based on their carbon intensity and fuel flexibility. For instance, a $50 per ton tax that rises 5% annually could cut up to 45% of U.S. emissions by 2040 (Macaluso et al. 16). While the power sector (25% of emissions) achieves almost 77% of the emissions reduction, the transportation sector (28% of emissions) is less responsive due to its heavy reliance on petroleum ("Total U.S."; Macaluso et al. 16; Hayes and Hafstead). In contrast, the industrial sector (23% of emissions) shows varied responses and risks emissions leakage to regions without carbon pricing ("Total U.S."; Hayes and Hafstead). Still, the residential and commercial sectors (13% of emissions) benefit indirectly from cleaner electricity and energy-efficient technologies ("Total U.S."; Hayes and Hafstead). Agriculture, not directly taxed, might see increased global emissions due to increased production cost and land use changes (Dumortier and Elobeid 1). Yet, with agriculture contributing only 10% to U.S. emissions in 2022, this effect remains minor compared to the potential 1.55 billion metric tons of CO2 reduction annually from a $10 carbon tax ("Total U.S."; Carroll and Stevens 1).

Carbon taxes can significantly reduce co-pollutants, improving public health. The Energy Innovation and Carbon Dividend Act (EICDA) could cut U.S. emissions from 4.74 billion metric tons in 2020 to 2.29 billion by 2035 (Hafstead "Annual Emissions"). EICDA would also lower local air pollutants, including a nearly 19% reduction in nitrogen oxides and almost 60% in sulfur oxides by 2030 (Hafstead "Local Air"). Reduced pollution, particularly PM2.5, lowers premature deaths, asthma, cardiovascular issues, and may reduce rates of autism and Alzheimer's (Ambasta and Buonocore 1). Economically, damages from PM2.5 are estimated at $500 billion to $1 trillion, but a carbon tax starting at $15 and rising to $30 per ton could save $92 billion (Muller). The Eastern U.S., especially the Ohio River Valley where communities live near coal and natural gas plants, could see substantial health benefits from improved air quality by 2030 (Muller). Thus, the public health and environmental benefits from carbon taxes highlight the importance for stringent emissions reductions.

Many carbon pricing systems let high-emission industries buy offsets, but these often fail to deliver real reductions, especially in sensitive areas like the Amazon and California's forests (Zakrzewski "A Carbon"). California's forest offset program, meant to cut emissions by preserving trees, has been criticized for generating up to 39 million "ghost credits" that do not reflect actual reductions (Temple and Songarchive). Developers exploit rules to gain credits from high-carbon areas, leading to financial gains without real climate benefits (Temple and Songarchive). CarbonPlan estimates that nearly one-third of these credits are overestimated, increasing emissions instead (Temple and Songarchive). California's new law AB 1305, effective January 1, 2024, requires businesses to disclose carbon offset information and imposes penalties up to $500,000 for non-compliance (McCullough et al.). This step toward stricter regulation underscores the need for a federal carbon tax to balance long-term emissions reductions with corporate flexibility.

4B. U.S. Competitiveness

Carbon taxes can reduce emissions but may also impact domestic industries and cause carbon leakage in energy-intensive sectors (Macaluso et al. 2). To mitigate these issues, the EU introduced Carbon Border Adjustment Mechanisms (CBAMs) in 2023, which impose fees on imports from high-emission countries to promote cleaner production and protect local industries (Zakrzewski "A Carbon"). The U.S. considered a similar CBAM proposal backed by Senators Cassidy and Cramer, which did not pass (United States, Congress, Senate 1). Research indicates that a $15 per ton CO2 price could reduce production by up to 5% in energy-intensive sectors, but the overall economic impact would be modest—around 1%—with only a slight rise in net imports (Aldy and Pizer 566, 590). Studies on the EU ETS show minimal evidence of significant leakage or competitiveness issues, suggesting concerns may be exaggerated (Zachmann and McWilliams 4). Implementing broad CBAMs could be complex and face WTO conflicts, while targeted CBAMs might lead to economic and job losses, similar to the impacts of tariffs on aluminum and steel (Zachmann and McWilliams 8). Instead of CBAMs, using carbon tax

revenues for monitoring, phased tax rates, and investing in clean technologies could better protect competitiveness and support the green transition.

4C. Economic Growth

An analysis of 31 EU countries, including 15 with additional carbon taxes, found neutral to positive effects on GDP and employment, particularly when revenues reduce other taxes (Stock and Metcalf 23-24). In British Columbia, a 2008 revenue-neutral carbon tax had minimal economic impact and increased growth by shifting jobs to less carbon-intensive sectors (Muresianu; Metcalf 31-32). Conversely, Australia's high carbon tax, without broader reforms, had negative effects and was quickly repealed (Muresianu). Evidence from regions like RGGI states and California suggests that carbon pricing does not impede GDP growth, implying similar results for a U.S. federal carbon tax. A $25-per-ton tax rising 5% annually could reduce the U.S. deficit by $865 billion from 2023 to 2032 (Pomerleau). The Energy Innovation and Carbon Dividend Act (EICDA) would increase inflation by only about 0.24% annually, well below the Federal Reserve's 2% target (Nuccitelli). While a $20-per-ton carbon fee might raise energy costs by $4 to $8 per month, rebates would generally offset these costs (Nuccitelli). Data from British Columbia show that carbon pricing doesn't necessarily cause inflation; revenue redistribution benefits lower-income households and taxes reduce fossil fuel demand (Nuccitelli). Additionally, Nordic countries investing in renewables have seen stable electricity prices, suggesting that a shift to renewables can mitigate long-term inflation risks (Nuccitelli). Overall, well-designed carbon pricing and revenue recycling can support economic growth while minimizing inflation.

However, general macroeconomic analyses often overlook sector-specific effects and the relative inelasticity of human capital in fossil fuel industries. A 2020 study on a $27 per ton U.S. carbon tax found it reduces total energy demand by 27%, with a 2.5% decrease in the dirty sector and a 0.5% increase in the clean sector, resulting in a ~3.6% drop in overall output (Fernández Intriago 48-49). Higher production costs in the dirty sector lower competitiveness, reducing employment by ~2.4% and increasing clean sector jobs by ~1% (Fernández Intriago 8). A separate study on a $35 per ton tax shows a modest overall unemployment rise (0.2 to 0.4%) but a 24% increase in coal sector unemployment, highlighting labor mobility constraints (Castellanos and Heutel 1). Shifts from high-skilled to low-skilled jobs in the dirty sector cause skill erosion and longer unemployment (Fernández Intriago 8). To mitigate job loss, carbon tax revenue can fund lump-sum rebates and labor tax cuts (Fernández Intriago 9; Castellanos and Heutel 31). Command-and-control policies may lower unemployment but are less efficient (Castellanos and Heutel 5). Therefore, a carbon tax should address labor market frictions and use targeted revenue recycling to minimize job losses and meet environmental goals effectively.

4D. Distributional Impact

Household carbon footprints in the U.S. differ significantly across income levels, locations, and urban settings. On average, households emit 24.2 tons of CO2 annually, with

lower-income households emitting 18.1 tons and higher-income households emitting 29.1 tons (Knittel and Green). Rural households have the highest average emissions at 27.7 tons, followed by suburban households at 26.0 tons and urban households at 21.1 tons (Knittel and Green). Geographic factors, such as the carbon intensity of electricity grids in regions like the Midwest, contribute to these disparities (Knittel and Green). Consequently, the financial burden of a carbon tax would vary widely, with households in regions dependent on fossil fuels or with extreme climates facing higher costs compared to those in areas with cleaner energy sources or milder weather (Ummel). Demographic factors also shape the impact of carbon taxes, which tend to be regressive, disproportionately affecting lower-income households that spend more on carbon-intensive goods (Fremstad and Paul). Without revenue recycling, a carbon tax would cost lower-income households about $907 annually (Knittel and Green).

Revenue recycling, particularly through carbon dividends, effectively mitigates the regressive effects of carbon taxes. A carbon tax with equal dividends becomes progressive, with 96% of households in the lowest income quintile seeing a net benefit of $307 per year, while higher-income households face net costs (Knittel and Green). Under the Energy Innovation and Carbon Dividend Act (EICDA), 54% of all households—and 92% of those in the lowest consumption quintile—benefit when the carbon tax is fully passed through to consumer prices; these figures rise to 67% overall and 98% for the lowest quintile with a 70% pass-through rate (Ummel). This redistribution shifts the burden to wealthier households with higher carbon footprints ("2020 Household"). Additional characteristics such as lower spending, larger family sizes, and greater reliance on public transportation can result in financial gains for households of color ("2020 Household"). Older and younger, larger households tend to benefit more from carbon dividends, which help offset these regressive impacts ("2020 Household").

Alternative revenue recycling methods, such as using carbon tax revenues to lower labor taxes or social security benefits, fail to protect low-income households adequately. Accordingly, the theoretical economic benefits from reduced labor taxes do not trickle down sufficiently to offset the increased costs of the carbon tax (Fremstad and Paul). Adjusting rebates for factors like appliance efficiency or local climate dilutes also the price signals needed to decrease energy consumption (Cronin et al.). Proposals for one-time payments based on household durable goods might reward past inefficiencies and penalize proactive choices (Cronin et al.). Adjusting rebates based on average consumption further minimizes inequities without undermining conservation incentives (Cronin et al.). Overall, household dividends effectively counter the regressive nature of carbon taxes by reducing inequalities, preventing excessive redistribution among similar-income groups, and aiding disadvantaged communities (Fremstad and Paul). Policymakers should adopt these strategies to ensure fair climate action, supporting vulnerable populations while advancing environmental goals.

4E. Revenue Recycling

Revenue recycling through direct rebates, green investments, and social policies can address financial concerns and garner political support. Directly rebating carbon tax revenues to

households has been effective in building public backing (Marshall). For instance, the "Conservative Case for Carbon Dividends" proposes a $40 per ton carbon fee with equal rebates to households, offering about $2,000 annually to a family of four, aligning with conservative principles while including border adjustments to protect U.S. competitiveness and encourage global adoption (Baker et al. 3). Studies from Germany and the U.S. suggest that revenue recycling could secure majority support for carbon taxes between $50 and $70 per ton if implemented internationally (Beiser-McGrath and Bernauer 1). Additionally, allocating revenues to renewable energy projects, energy efficiency, or subsidies for low-carbon technologies like electric vehicles bolsters the green sector, countering fossil fuel and industry opposition (Beiser-McGrath and Bernauer 6). Notably, 79% of Americans prioritize renewable energy development over expanding fossil fuels (Tyson and Kennedy). Integrating climate policies with social reforms, such as just transition assistance, affordable housing, or a $15 minimum wage, can enhance support, especially in fossil fuel and marginalized communities (Gazmararian 2; Bergquist et al. 2). These strategies can help foster broader climate action coalitions.

Even well-designed policies need effective political messaging to gain support. Educational efforts can increase public approval—explaining a $50 carbon fee raised U.S. support from 58% to 70%—but negative political messaging and polarization often undermine these gains (Marshall "How to"). Political narratives can undermine the effectiveness of rebates and contribute to widespread misunderstandings about carbon taxes, with only 15% believing they would significantly reduce pollution (Fremstad et al. 5-6; Marshall "Building" 10). Moreover, industry lobbying and negative advertising further diminish public support (Levi et al. 132-133; Marshall "Building" 3-4). Strategic framing, including terms like "carbon fee" or "carbon cashback" rather than "tax," can improve policy acceptance (Marshall "Building" 15; "Carbon Cashback"; Karapin 143). Clearly communicating benefits, managing price increases, and addressing fairness are crucial for building support (Marshall "Building" 6-7). Successful carbon pricing depends on a supportive political environment, effective revenue recycling, and strategic messaging.

4F. Literature Review Synthesis

The literature suggests that a successful federal carbon tax in the U.S. must balance emissions reductions, economic impacts, competitiveness, distributional effects, and political feasibility. By internalizing CO2 damages, a carbon tax can lower emissions and improve public health, though issues like the integrity of carbon offsets, greenwashing, and carbon leakage necessitate robust regulation. Economically, a carbon tax might raise energy costs and affect consumer prices, potentially slowing growth and impacting U.S. competitiveness if other countries do not adopt similar measures. To address competitiveness, especially in energy-intensive industries, Carbon Border Adjustment Mechanisms (CBAMs) could help but involve complexities such as WTO concerns and market distortions. Advancing clean energy investments and decarbonizing sectors can enhance energy independence and protect domestic industries. While a carbon tax might increase the cost of goods, studies suggest minimal impacts

on GDP and employment, especially if revenues are recycled through direct dividends to households, which can offset regressive effects on lower-income and fossil-dependent groups. Revenue recycling and clean energy transitions can also unite diverse stakeholders, including climate advocates, environmental justice groups, fossil fuel communities, green companies, and conservative leaders. Political success relies on strategic communication and stakeholder engagement, with phased implementation and clear messaging, such as framing the tax as "carbon cashback," essential for broader acceptance. In summary, a well-designed federal carbon tax that balances emissions reductions, economic impacts, and equity can build strong political support across diverse stakeholder groups.

5. Conclusion

5A. Key Takeaways

Grounded in the WHO, HOW, WHAT EJ framework proposal, stakeholder mapping (see fig. 1), literature reviews, and case studies, this research offers criteria for building diverse coalitions to promote an economically effective, politically feasible, and environmentally just carbon tax policy. A stringent carbon tax should cover diverse emissions sources, increase rates over time, and include robust regulations to ensure compliance and equitable pollution reductions in environmental justice communities. Carbon tax revenues should fund clean technology, support environmental justice infrastructure, aid fossil fuel-dependent communities in transition, and provide household dividends to offset costs and build public support. Actively involving stakeholders in decision-making, ensuring transparency about policy benefits and revenue use, and incorporating public feedback can build trust and foster equitable, responsive climate policies. Overall, a well-designed federal carbon tax can harmonize diverse stakeholders, balancing justice, efficiency, and feasibility.

5B. Limitations and Future Research

Longitudinal studies are needed to evaluate the long-term effects of carbon taxes on health, pollution reduction, and equity, expanding the timeframe of political economy considerations. This paper focuses on carbon pricing through the SCC and carbon taxes but expanding stakeholder analysis to include other climate policies—such as regulation standards, renewable energy subsidies, and fossil fuel subsidy reform—is crucial for advancing a green transition. Future research should explore the local and global impacts of federal carbon taxes, their interaction with existing mechanisms, and the legal challenges in climate policy enforcement. Additionally, integrating circular economy theories can enhance climate policy effectiveness and provide new analytical frameworks. Scenario planning for unforeseen factors like economic crises, technological advancements, and political shifts will support the development of resilient climate policies. Addressing these areas will contribute to a comprehensive understanding of effective and inclusive climate strategies.

Works Referenced

"Additional Statistics on Energy." Government of Canada, natural-resources.canada.ca/maps-tools-and-publications/publications/energy-publications/energy-efficiency-publications/additional-statistics-on-energy/1239. Accessed 27 Aug. 2024.

Aldy, Joe, et al. How Is the US Pricing Carbon? How Could We Price Carbon? Washington, DC, Resources for the Future, 13 Oct. 2022, media.rff.org/documents/WP_22-19_GN4gxYW.pdf. Accessed 29 July 2024.

Aldy, Joseph E., and William A. Pizer. "The Competitiveness Impacts of Climate Change Mitigation Policies." Journal of the Association of Environmental and Resource Economists, vol. 2, no. 4, Dec. 2015, pp. 565-95, https://doi.org/10.1086/683305. Accessed 27 Aug. 2024.

Aldy, Joseph E., and Robert N. Stavins. "The Promise and Problems of Pricing Carbon." The Journal of Environment & Development, vol. 21, no. 2, 18 Apr. 2012, pp. 152-80, https://doi.org/10.1177/1070496512442508. Accessed 2 Sept. 2024.

Ambasta, Anshula, and Jonathan J. Buonocore. "Carbon Pricing: A Win-win Environmental and Public Health Policy." Canadian Journal of Public Health, vol. 109, nos. 5-6, 28 June 2018, pp. 779-81, https://doi.org/10.17269/s41997-018-0099-5. Accessed 28 Aug. 2024.

"ARB Emissions Trading Program." Air Resources Board, 9 Feb. 2015, ww2.arb.ca.gov/sites/default/files/cap-and-trade/guidance/cap_trade_overview.pdf. Accessed 26 Aug. 2024.

Aronoff, Kate. "BP's Suspicious Support for a Carbon Market in Washington State." The New Republic, 6 May 2021, newrepublic.com/article/162313/bp-carbon-offsets-washington-finite-carbon-carlyle. Accessed 27 Aug. 2024.

Backman, Isabella. "Stanford Explainer: Social Cost of Carbon." Stanford Report, Stanford University, 17 June 2021, news.stanford.edu/stories/2021/06/professors-explain-social-cost-carbon. Accessed 8 June 2024.

Baker, James A., et al. The Conservative Case for Carbon Dividends. Climate Leadership Council, Feb. 2017, clcouncil.org/media/2017/03/The-Conservative-Case-for-Carbon-Dividends.pdf. Accessed 3 Sept. 2024.

Bakken, Rebecca. "What Is Sustainable Finance and Why Is It Important?" Harvard Extension School Blog, 9 Aug. 2021, extension.harvard.edu/blog/what-is-sustainable-finance-and-why-is-it-important/. Accessed 19 Aug. 2024.

Baranzini, Andrea, et al. "Carbon Pricing in Climate Policy: Seven Reasons, Complementary Instruments, and Political Economy Considerations." *WIREs Climate Change*, vol. 8, no. 4, 31 Mar. 2017, https://doi.org/10.1002/wcc.462. Accessed 3 Sept. 2024.

Barbanell, Melissa. "A Brief Summary of the Climate and Energy Provisions of the Inflation Reduction Act of 2022." *World Resources Institute*, 28 Oct. 2022, www.wri.org/update/brief-summary-climate-and-energy-provisions-inflation-reduction-act-2022. Accessed 2 Sept. 2024.

Barrage, Lint, and William Nordhaus. "Policies, Projections, and the Social Cost of Carbon: Results from the DICE-2023 Model." *Proceedings of the National Academy of Sciences*, vol. 121, no. 13, 19 Mar. 2024, https://doi.org/10.1073/pnas.2312030121. Accessed 14 June 2024.

Beiser-McGrath, Liam F., and Thomas Bernauer. "Could Revenue Recycling Make Effective Carbon Taxation Politically Feasible?" *Science Advances*, vol. 5, no. 9, 6 Sept. 2019, https://doi.org/10.1126/sciadv.aax3323. Accessed 3 Sept. 2024.

Bellora, Cecilia, and Lionel Fontagné. "EU in Search of a Carbon Border Adjustment Mechanism." *Energy Economics*, vol. 123, July 2023, p. 106673, https://doi.org/10.1016/j.eneco.2023.106673. Accessed 22 Aug. 2024.

Bergquist, Parrish, et al. "Combining Climate, Economic, and Social Policy Builds Public Support for Climate Action in the US." *Environmental Research Letters*, vol. 15, no. 5, 1 May 2020, p. 054019, https://doi.org/10.1088/1748-9326/ab81c1. Accessed 3 Sept. 2024.

Bernton, Hal. "As Washington Debates Carbon Fee, One Oil Giant Is Opposed but Another Is Silent; What's That About?" *The Seattle Times*, 30 Sept. 2018, www.seattletimes.com/seattle-news/politics/shell-bp-go-separate-ways-as-washington-voters-consider-fee-on-greenhouse-gas-polluters/. Accessed 27 Aug. 2024.

Bill Cassidy, M.D. "Cassidy Leads Republican Senate Opposition to a Carbon Tax." 26 Oct. 2023, www.cassidy.senate.gov/newsroom/press-releases/cassidy-leads-republican-senate-opposition-to-a-carbon-tax/. Accessed 20 Aug. 2024.

Black, Simon, et al. "More Countries Are Pricing Carbon, but Emissions Are Still Too Cheap." *International Monetary Fund*, 1 July 2022, www.imf.org/en/Blogs/Articles/2022/07/21/blog-more-countries-are-pricing-carbon-but-emissions-are-still-too-cheap#:~:text=To%20limit%20global%20warming%2C%20coverage,an%20IMF%20Staff%20Climate%20Note. Accessed 29 July 2024.

Bonasia, Christopher. "Ottawa Faces 'Profound' Political Impacts after Carbon Tax Messaging Falls Flat." *The Energy Mix*, 20 Feb. 2023, www.theenergymix.com/ottawa-faces-profound-political-impacts-after-carbon-tax-messaging-falls-flat/. Accessed 24 Aug. 2024.

Borenstein, Severin, and Ryan Kellogg. *Carbon Pricing, Clean Electricity Standards, and Clean Electricity Subsidies on the Path to Zero Emissions*. National Bureau of Economic Research, July 2022, https://doi.org/10.3386/w30263. Accessed 3 Sept. 2024.

Boyce, James K., et al. "Environmental Justice and Carbon Pricing: Can They Be Reconciled?" *Global Challenges*, vol. 7, no. 4, 28 Feb. 2023. *Wiley Online Library*, https://doi.org/10.1002/gch2.202200204. Accessed 20 Aug. 2024.

"California Cap and Trade." *Center for Climate and Energy Solutions*, www.c2es.org/content/california-cap-and-trade/. Accessed 26 Aug. 2024.

"Canada Federal Output-Based Pricing System." *International Carbon Action Partnership*, icapcarbonaction.com/system/files/ets_pdfs/icap-etsmap-factsheet-135.pdf. Accessed 27 Aug. 2024.

"Cap-and-invest Linkage." *Department of Ecology State of Washington*, ecology.wa.gov/air-climate/climate-commitment-act/cap-and-invest/linkage. Accessed 26 Aug. 2024.

"Cap and Trade Vs. Taxes." *Center for Climate and Energy Solutions*, Mar. 2009, www.c2es.org/document/cap-and-trade-vs-taxes/. Accessed 2 Sept. 2024.

"Carbon Cashback Handouts." *CCL Community*, community.citizensclimate.org/resources/item/19/519. Accessed 3 Sept. 2024.

"Carbon Pollution Pricing Systems across Canada." *Government of Canada*, www.canada.ca/en/environment-climate-change/services/climate-change/pricing-pollution-how-it-will-work.html. Accessed 27 Aug. 2024.

"Carbon Pricing Bills in Congress." *CCL Community*, community.citizensclimate.org/resources/item/19/220#heading_0. Accessed 2 Sept. 2024.

"Carbon Pricing Is a False Solution to Climate Chaos." *Indigenous Environmental Network*, www.ienearth.org/carbon-pricing/. Accessed 11 May 2023.

"Carbon Tax Basics." *Center for Climate and Energy Solutions*, www.c2es.org/content/carbon-tax-basics/#:~:text=Competitiveness%20%E2%80%93%20Without%20provisions%20protecting%20local%20production%2C,that%20do%20not%20face%20an%20equivalent%20price. Accessed 18 Aug. 2024.

Carl, Jeremy, and David Fedor. "Tracking Global Carbon Revenues: A Survey of Carbon Taxes versus Cap-and-trade in the Real World." *Energy Policy*, vol. 96, Sept. 2016, pp. 50-77, https://doi.org/10.1016/j.enpol.2016.05.023. Accessed 2 Sept. 2024.

Carleton, Tamma, and Michael Greenstone. "A Guide to Updating the US Government's Social Cost of Carbon." *Review of Environmental Economics and Policy*, vol. 16, no. 2, 1 June 2022, pp. 196-218, https://doi.org/10.1086/720988.

Carroll, Deborah A., and Kelly A. Stevens. "The Short-term Impact on Emissions and Federal Tax Revenue of a Carbon Tax in the U.S. Electricity Sector." *Energy Policy*, vol. 158, Nov. 2021, p. 112526, https://doi.org/10.1016/j.enpol.2021.112526. Accessed 28 Aug. 2024.

Casey, Gregory, et al. "The Macroeconomics of Clean Energy Subsidies." *SSRN Electronic Journal*, 2023, https://doi.org/10.2139/ssrn.4665589. Accessed 3 Sept. 2024.

Castellanos, Kenneth, and Garth Heutel. *Unemployment, Labor Mobility, and Climate Policy*. National Bureau of Economic Research, May 2019, https://doi.org/10.3386/w25797. Accessed 27 Aug. 2024.

Cerrell Associates. *Political Difficulties Facing Waste-to-Energy Conversion Plant Siting*. 1984, www.ejnet.org/ej/cerrell.pdf. Accessed 30 Aug. 2024.

Chen, Qian, et al. "The Influence of Carbon Tax on CO2 Rebound Effect and Welfare in Chinese Households." *Energy Policy*, vol. 168, Sept. 2022, p. 113103, https://doi.org/10.1016/j.enpol.2022.113103. Accessed 27 Aug. 2024.

Chris Fawsona, et al. *Carbon Pricing in the US Private Sector*. Center for Growth and Opportunity, Mar. 2019, www.thecgo.org/wp-content/uploads/2020/10/Carbon-Pricing-in-the-US-Private-Sector.pdf. Accessed 18 Aug. 2024.

"Climate Action Framework." *American Petroleum Institute*, www.api.org/climate#carbon-price. Accessed 18 Aug. 2024.

"Climate Justice." *Washington Department of Health*, doh.wa.gov/community-and-environment/climate-and-health/climate-justice#:~:text=An%20important%20distinction%20between%20environmental,and%20recovering%20from%20climate%20events. Accessed 30 Aug. 2024.

Climate Justice Advocates. "Concerns with SB 5126, Cap and Trade." Letter to Honorable State Senators and State Representatives, 1 Apr. 2021, drive.google.com/file/d/1MhtbbzSTKhNsva_heyQS2sS6pPu7MSH5/view. Accessed 26 Aug. 2024.

"The Climate Leadership Council's Carbon Dividends Plan." *CCL Community*, community.citizensclimate.org/resources/item/19/136#heading_0. Accessed 3 Sept. 2024.

"CO2 Emissions per Capita." *Worldometer*, www.worldometers.info/co2-emissions/co2-emissions-per-capita/#google_vignette. Accessed 27 Aug. 2024.

Coletto, David. "Understanding Canadian Perceptions of the Climate Action Incentive Payment and the Carbon Tax: An In-Depth Poll Analysis." *Abacus Data*, 30 Jan. 2024, abacusdata.ca/carbon-tax-pollution-pricing-carbon-action-incentive-payment-abacus-data-polling/. Accessed 24 Aug. 2024.

"Cosponsors: H.R.2307 — 117th Congress (2021-2022)." *Congress.gov*, www.congress.gov/bill/117th-congress/house-bill/2307#:~:text=Energy%20Innovation%20and%20Carbon%20Dividend%20Act%20of%202021,-This%20bill%20imposes&text=The%20rate%20begins%20at%20%2415,meeting%20specified%20emissions%20reduction%20targets. Accessed 2 Sept. 2024.

Council of Economic Advisers. "Valuing the Future: Revision to the Social Discount Rate Means Appropriately Assessing Benefits and Costs." *The White House*, 27 Feb. 2024, www.whitehouse.gov/cea/written-materials/2024/02/27/valuing-the-future-revision-to-the-social-discount-rate-means-appropriately-assessing-benefits-and-costs/. Accessed 18 July 2024.

Cronin, Julie Anne, et al. "Vertical and Horizontal Redistributions from a Carbon Tax and Rebate." *Journal of the Association of Environmental and Resource Economists*, vol. 6, no. 1, Mar. 2019, www.nber.org/papers/w23250. Accessed 28 Aug. 2024.

Declet-Barreto, Juan, and Andrew A. Rosenberg. "Environmental Justice and Power Plant Emissions in the Regional Greenhouse Gas Initiative States." *PLOS One*, vol. 17, no. 7, 20 July 2022, p. e0271026, https://doi.org/10.1371/journal.pone.0271026. Accessed 26 Aug. 2024.

"Development of EU ETS (2005-2020)." *European Comission*, climate.ec.europa.eu/eu-action/eu-emissions-trading-system-eu-ets/development-eu-ets-2005-2020_en#:~:text=Set%20up%20in%202005%2C%20the,phase%20(2021%2D2030). Accessed 24 Aug. 2024.

Diaz, Shea. "Getting to the Root of Environmental Injustice." *Environmental Law Journal*, New York University School of Law, 20 Jan. 2016, nyuelj.org/2016/01/getting-to-the-root-of-environmental-injustice/. Accessed 30 Aug. 2024.

Diedrich, Graham. "Carbon Pricing: Carbon Markets and Carbon Taxes." *Michigan State University*, 24 Jan. 2022, www.canr.msu.edu/news/overview-carbon-pricing-carbon-markets-and-carbon-taxes. Accessed 2 Sept. 2024.

Doniger, Alicia. "As Climate Change Policy Takes Shape, Will the U.S. Ever Put a Price on Carbon?" *CNBC*, 15 Nov. 2021, www.cnbc.com/2021/11/15/will-us-ever-put-a-price-on-carbon-as-part-of-climate-change-policy.html#:~:text=Historically%2C%20there%20has%20been%20some,reconciliation%20process%2C%E2%80%9D%20Newell%20said. Accessed 24 Aug. 2024.

Downie, Christian, and Robert Brulle. "Big Oil's Allies Spend Big Money on Ads and Lobbying to Keep Fossil Fuels Flowing." *Pennsylvania Capital-Star*, 14 Feb. 2023, penncapital-star.com/commentary/big-oils-allies-spend-big-money-on-ads-and-lobbying-to-keep-fossil-fuels-flowing-analysis/. Accessed 21 Aug. 2024.

Dumortier, Jerome, and Amani Elobeid. "Effects of a Carbon Tax in the United States on Agricultural Markets and Carbon Emissions from Land-use Change." *Land Use Policy*, vol. 103, Apr. 2021, p. 105320, https://doi.org/10.1016/j.landusepol.2021.105320. Accessed 28 Aug. 2024.

---. "Implications of a US Carbon Tax on Agricultural Markets and GHG Emissions from Land-use Change." *Agricultural Policy Review*, winter 2020. *Center for Agricultural and Rural Development, Iowa State University*, www.card.iastate.edu/ag_policy_review/article/?a=106. Accessed 28 Aug. 2024.

"Elements of RGGI." *The Regional Greenhouse Gas Initiative*, www.rggi.org/program-overview-and-design/elements. Accessed 26 Aug. 2024.

"Energy Innovation Act Progress Page." *CCL Community*, community.citizensclimate.org/resources/item/19/401. Accessed 2 Sept. 2024.

"Environmental Justice." *United States Environmental Protection Agency*, www.epa.gov/environmentaljustice. Accessed 30 Aug. 2024.

"Environmental Justice: Addressing the Burden of Air Pollution." *American Lung Assosiation*, 26 Aug. 2020, www.lung.org/blog/environmental-justice-air-pollution#:~:text=In%20fact%2C%20people%20of%20color,it%27s%20about%20health%20and%20lives. Accessed 30 Aug. 2024.

EPA Report on the Social Cost of Greenhouse Gases: Estimates Incorporating Recent Scientific Advances. Research report no. EPA-HQ-OAR-2021-0317, U.S. Environmental Protection Agency, Nov. 2023. *EPA*, www.epa.gov/system/files/documents/2023-12/epa_scghg_2023_report_final.pdf. Accessed 15 July 2024.

Estien, Cesar O., et al. "Historical Redlining Is Associated with Disparities in Environmental Quality across California." *Environmental Science & Technology Letters*, vol. 11, no. 2, 19 Jan. 2024, pp. 54-59, https://doi.org/10.1021/acs.estlett.3c00870. Accessed 30 Aug. 2024.

Evans, Simon, et al. "Q&A: The Social Cost of Carbon." *Carbon Brief*, 14 Feb. 2017, www.carbonbrief.org/qa-social-cost-carbon/. Accessed 25 June 2024.

Fairbrother, Malcolm, and Ekaterina Rhodes. "Climate Policy in British Columbia: An Unexpected Journey." *Frontiers in Climate*, vol. 4, 18 Jan. 2023, https://doi.org/10.3389/fclim.2022.1043672. Accessed 24 Aug. 2024.

"FAQ Cap-and-Trade Program." *California Air Resources Board*, ww2.arb.ca.gov/resources/documents/faq-cap-and-trade-program. Accessed 26 Aug. 2024.

Farber, Daniel. "State Governmental Leadership in U.S. Climate Policy." *Wilson Center*, 23 June 2021, www.wilsoncenter.org/article/state-governmental-leadership-us-climate-policy. Accessed 20 Aug. 2024.

Federal Register. National Archives and Records Administration, www.federalregister.gov/documents/search?conditions%5Bterm%5D=%22social+cost+of+carbon%22. Accessed 15 July 2024.

Federal Register. www.federalregister.gov/documents/search?conditions%5Bterm%5D=%22social+cost+of+carbon%22. Accessed 3 Sept. 2024.

Fernández Intriago, Luis A. *Carbon Taxation, Green Jobs, and Sectoral Human Capital*. 31 Jan. 2020. *ETH Zürich*, ethz.ch/content/dam/ethz/special-interest/mtec/cer-eth/resource-econ-dam/documents/research/sured/sured-2020/Carbon%20Taxation,%20Green%20Jobs,%20and%20Sectoral%20Human%20Capital.pdf. Accessed 27 Aug. 2024.

Fifth National Climate Assessment: Report-in-Brief. U.S. Global Change Research Program, 2023, https://doi.org/10.7930/nca5.2023.rib. Accessed 2 Sept. 2024.

Fremstad, Anders, et al. "The Role of Rebates in Public Support for Carbon Taxes." *Environmental Research Letters*, vol. 17, no. 8, 1 Aug. 2022, p. 084040, https://doi.org/10.1088/1748-9326/ac8607. Accessed 30 July 2024.

Fremstad, Anders, and Mark Paul. "The Impact of a Carbon Tax on Inequality." *Ecological Economics*, vol. 163, Sept. 2019, pp. 88-97, https://doi.org/10.1016/j.ecolecon.2019.04.016. Accessed 28 Aug. 2024.

Friedman, Lisa. "The Zombies of the U.S. Tax Code: Why Fossil Fuels Subsidies Seem Impossible to Kill." *The New York Times*, 20 Mar. 2024, www.nytimes.com/2024/03/15/climate/tax-breaks-oil-gas-us.html. Accessed 20 Aug. 2024.

Gazmararian, Alexander F. "Fossil Fuel Communities Support Climate Policy Coupled with Just Transition Assistance." *Energy Policy*, vol. 184, Jan. 2024, p. 113880, https://doi.org/10.1016/j.enpol.2023.113880. Accessed 22 Aug. 2024.

"GDP per Capita." *Worldometer*, 2022, ourworldindata.org/grapher/gdp-per-capita-worldbank?tab=table. Accessed 24 Aug. 2024. Map.

"Gini Coefficient, 2022." *Our World in Data*, ourworldindata.org/economic-inequality. Accessed 24 Aug. 2024. Map.

Golden, Hallie. "Effort to Repeal Washington's Carbon Program Puts Budget in Limbo with Billions at Stake." *The Seattle Times*, 28 Feb. 2024, www.seattletimes.com/seattle-news/effort-to-repeal-washingtons-landmark-carbon-program-puts-budget-in-limbo-with-billions-at-stake/. Accessed 27 Aug. 2024.

Governor Gavin Newsom. "California's Cap-and-Trade Program Funds $28 Billion in Climate Investments over Last Decade." 28 May 2024,

www.gov.ca.gov/2024/05/08/californias-cap-and-trade-program-funds-28-billion-in-climate-investments-over-last-decade/. Accessed 26 Aug. 2024.

"'Green Future Act' Corporate Polluter Fee." *Our Climate*, docs.google.com/document/d/1jxg24mvvr2Q0JOFo8E-we_Dm7hWiqfjOEmStquGJXNA/edit. Accessed 13 May 2023.

"Green Future Now's Public Statement on Equity and Environmental Justice." *Our Climate*, docs.google.com/document/d/1eEnXTC9iMPwkREbsaDWsf8hzO4a53-1PL7HmeUVqLwU/edit. Accessed 13 May 2023.

Hafstead, Marc. Annual Emissions: Carbon Pricing Calculator. *Resources for the Future*, 10 Aug. 2020, www.rff.org/publications/data-tools/carbon-pricing-calculator/. Accessed 27 Aug. 2024. Chart.

---. Annual Revenues: Carbon Pricing Calculator. *Resources for the Future*, 10 Aug. 2020, www.rff.org/publications/data-tools/carbon-pricing-calculator/. Accessed 27 Aug. 2024. Chart.

---. "Carbon Pricing 101." *Resources for the Future*, 3 Mar. 2022, www.rff.org/publications/explainers/carbon-pricing-101/. Accessed 3 Sept. 2024.

---. Consumer Prices: Carbon Pricing Calculator. *Resources for the Future*, 10 Aug. 2020, www.rff.org/publications/data-tools/carbon-pricing-calculator/. Accessed 27 Aug. 2024. Chart.

---. Local Air Pollutants: Carbon Pricing Calculator. *Resources for the Future*, 10 Aug. 2020, www.rff.org/publications/data-tools/carbon-pricing-calculator/. Accessed 27 Aug. 2024. Chart.

Hart Research Associates, and Chesapeake Beach Consulting. "Findings from a Survey in States Participating in RGGI." 9 Aug. 2016, www.sierraclub.org/sites/default/files/program/documents/FOR%20RELEASE%20RGGI%20Survey%202016%20Memo.pdf. Accessed 26 Aug. 2024.

Hayes, Kristin, and Marc Hafstead. "Carbon Pricing 103: Effects across Sectors." *Resources for the Future*, 27 Apr. 2020, media.rff.org/documents/Carbon_Pricing_103.pdf. Accessed 27 Aug. 2024.

Hiller, Jennifer, and Svea Herbst-Bayliss. "Exxon Loses Board Seats to Activist Hedge Fund in Landmark Climate Vote." *Reuters*, 26 May 2021, www.reuters.com/business/sustainable-business/shareholder-activism-reaches-milestone-exxon-board-vote-nears-end-2021-05-26/. Accessed 20 Aug. 2024.

Hines-Acosta, Lauren. "RGGI Officially Removed from Virginia State Budget." *Bay Journal*, 22 May 2024, www.bayjournal.com/news/policy/rggi-officially-removed-from-virginia-state-budget/article_2e69630a-1897-11ef-addc-abcb1aae1b72.html. Accessed 26 Aug. 2024.

Ho, Ben. "Prioritising Innovation: The Case against the Carbon Tax." *Reimagining Economics for a Carbon-constrained World*, Economist, 16 Nov. 2022,

impact.economist.com/sustainability/net-zero-and-energy/prioritising-innovation-the-case-against-the-carbon-tax-ben-ho. Accessed 12 Aug. 2024.

"The Investment of RGGI Proceeds in 2022." *The Regional Greenhouse Gas Initiative*, July 2024, www.rggi.org/sites/default/files/Uploads/Proceeds/RGGI_Proceeds_Report_2022.pdf. Accessed 26 Aug. 2024.

Isabell Sustainability Scholar Proposal. docs.google.com/document/d/1-gtoEv6VJFCfi8PG_nZrgeMjuEggAvo5r1Sje2onlVY/edit. Accessed 5 June 2024.

Jenkins, Jesse D. *Why Carbon Pricing Falls Short*. Kleinman Center for Energy.Policy, Apr. 2019, kleinmanenergy.upenn.edu/wp-content/uploads/2020/08/KCEP-Why-Carbon-Pricing-Falls-Short-Digest-singles-1.pdf. Accessed 2 Sept. 2024.

"Justice40." *The White House*, www.whitehouse.gov/environmentaljustice/justice40/. Accessed 30 Aug. 2024.

Karapin, Roger. "The Political Viability of Carbon Pricing: Policy Design and Framing in British Columbia and California." *Review of Policy Research*, vol. 37, no. 2, Mar. 2020, pp. 140-73, https://doi.org/10.1111/ropr.12373. Accessed 24 Aug. 2024.

Kaswan, Alice. "Environmental Justice and Environmental Law." *Fordham Environmental Law Review*, vol. 24, no. 2, spring 2013, pp. 149-79. *JSTOR*, www.jstor.org/stable/26195842. Accessed 30 Aug. 2024.

Kaufman, Noah. "Carbon Tax Vs. Cap-and-Trade: What's a Better Policy to Cut Emissions?" *World Resources Institute*, 1 Mar. 2016, www.wri.org/insights/carbon-tax-vs-cap-and-trade-whats-better-policy-cut-emissions. Accessed 2 Sept. 2024.

---. "What You Need to Know about a Federal Carbon Tax in the United States." *Center on Global Energy Policy*, 2 Apr. 2019, www.energypolicy.columbia.edu/publications/what-you-need-to-know-about-a-federal-carbon-tax-in-the-united-states/#:~:text=Several%20of%20the%20most%20important,the%20impacts%20on%20energy%20prices. Accessed 3 Sept. 2024.

Knittel, Christopher R. "Diary of a Wimpy Carbon Tax: Carbon Taxes as Federal Climate Policy." *MIT Center for Energy and Environmental Policy Research*, Aug. 2020, ceepr.mit.edu/wp-content/uploads/2021/09/2019-013.pdf. Accessed 22 Apr. 2023.

Knittel, Christopher R., and Thomas W. Green. *Distributed Effects of Climate Policy: A Machine Learning Approach*. MIT Center for Energy and Environmental Policy Research, 1 Sept. 2020. The Roosevelt Project Special Series. *JSTOR*, www.jstor.org/stable/resrep34668. Accessed 28 Aug. 2024.

Laeven, Luc, and Alexander Popov. "Carbon Taxes and the Geography of Fossil Lending." *Journal of International Economics*, vol. 144, Sept. 2023, p. 103797, https://doi.org/10.1016/j.jinteco.2023.103797. Accessed 20 Aug. 2024.

Larson, John, et al. *Energy and Environmental Implications of a Carbon Tax in the United States*. Edited by Noah Kaufman, New York City, Columbia SIPA Center on Global Energy Policy, July 2018, www.energypolicy.columbia.edu/sites/default/files/pictures/CGEP_Energy_Environmental_Impacts_CarbonTax_FINAL.pdf. Accessed 22 Aug. 2024.

Lavelle, Marianne. "How Big Oil Blocked Gov. Jay Inslee on Climate Change." *The Weather Channel*, features.weather.com/collateral/big-oil-blocked-gov-jay-inslee-climate-change/. Accessed 27 Aug. 2024.

Laville, Sandra. "Top Oil Firms Spending Millions Lobbying to Block Climate Change Policies, Says Report." *The Guardian*, www.theguardian.com/business/2019/mar/22/top-oil-firms-spending-millions-lobbying-to-block-climate-change-policies-says-report. Accessed 18 Aug. 2024.

Layden, Samantha. "Land of the Free? Environmental Racism and Its Impact on Cancer Alley, Louisiana." *Keele University*, www.keele.ac.uk/extinction/controversy/canceralley/. Accessed 30 Aug. 2024.

Lessmann, Christian, and Niklas Kramer. "The Effect of Cap-and-trade on Sectoral Emissions: Evidence from California." *Energy Policy*, vol. 188, May 2024, p. 114066, https://doi.org/10.1016/j.enpol.2024.114066. Accessed 26 Aug. 2024.

Levi, Sebastian, et al. "Political Economy Determinants of Carbon Pricing." *Global Environmental Politics*, vol. 20, no. 2, May 2020, pp. 128-56, https://doi.org/10.1162/glep_a_00549. Accessed 2 Sept. 2024.

Levinson, Arik, et al. "The Inflation Reduction Act's Benefits and Costs." *U.S. Department of the Treasury*, 1 Mar. 2024, home.treasury.gov/news/featured-stories/the-inflation-reduction-acts-benefits-and-costs. Accessed 3 Sept. 2024.

Lowenstein, Adam. "Meet the DC Thinktank Giving Big Oil 'the Opportunity to Say They've Done Something.'" *The Guardian*, 9 July 2023, www.theguardian.com/us-news/2023/jul/09/climate-leadership-council-big-oil-thinktank. Accessed 3 Sept. 2024.

Macaluso, Nick, et al. "The Impact of Carbon Taxation and Revenue Recycling on U.S. Industries." *Climate Change Economics*, vol. 09, no. 01, Feb. 2018, p. 1840005, https://doi.org/10.1142/s2010007818400055. Accessed 27 Aug. 2024.

Manion, Michelle, et al. "Analysis of the Public Health Impacts of the Regional Greenhouse Gas Initiative." *Abt Global*, 11 Jan. 2017, www.abtglobal.com/insights/publications/report/analysis-of-the-public-health-impacts-of-the-regional-greenhouse-gas. Accessed 26 Aug. 2024.

Marshall, Jonathan. "Building Support for Carbon Pricing: A Research Guide." *Citizens' Climate Lobby*, Mar. 2023, static.prod01.ue1.p.pcomm.net/cclobby/content/contents/training/Economics/Carbon-Tax-Political-Research-Guide.pdf. Accessed 30 July 2024.

---. *How "Carbon Fee and Dividend" Serves Economic and Environmental Justice*. drive.google.com/file/d/1Uot9y2aUbNP6qAdOZpOiJ9lsBP7KbFOd/view. Accessed 13 May 2023.

---. "How to Make Carbon Pricing More Popular." *Citizens' Climate Lobby*, 29 Jan. 2023, citizensclimatelobby.org/blog/policy/how-to-make-carbon-pricing-more-popular/. Accessed 30 July 2024.

Martinez, Chris, et al. *These Fossil Fuel Industry Tactics Are Fueling Democratic Backsliding*. Center for American Progress, 5 Dec. 2023, www.americanprogress.org/article/these-fossil-fuel-industry-tactics-are-fueling-democratic-backsliding/. Accessed 21 Aug. 2024.

McCullough,, David M., et al. "California Combats Greenwashing with New Voluntary Carbon Offset and Carbon-Neutral and Low-Carbon Product Disclosure Law." *Pillsbury*, 19 Dec. 2023, www.pillsburylaw.com/en/news-and-insights/california-greenwashing-carbon-law.html. Accessed 28 Aug. 2024.

McFarlane, Lydia. "Spending on Environmental Lobbying on the Rise During Biden Administration." *Investigate Midwest*, 16 Feb. 2024, investigatemidwest.org/2024/02/16/spending-on-environmental-lobbying-on-the-rise-during-biden-administration/. Accessed 3 Sept. 2024.

McGreal, Chris. "Big Oil and Gas Kept a Dirty Secret for Decades. Now They May Pay the Price." *The Guardian*, 3 june 2021, www.theguardian.com/environment/2021/jun/30/climate-crimes-oil-and-gas-environment. Accessed 18 Aug. 2024.

Mesa, Natalia. "Washington's Controversial Cap-and-Trade Program, Explained. Really." *HighCountry News*, 20 Mar. 2024, www.hcn.org/articles/washingtons-controversial-cap-and-trade-program-explained-really/. Accessed 26 Aug. 2024.

Metcalf, Gilbert E. *On the Economics of a Carbon Tax for the United States*. 8 Mar. 2019, www.brookings.edu/wp-content/uploads/2019/03/On-the-Economics-of-a-Carbon-Tax-for-the-United-States.pdf. Accessed 27 Aug. 2024.

Miron, Jeffrey. "Clean Energy Subsidies Vs. a Carbon Tax." *CATO at Liberty*, CATO Institute, 22 Jan. 2024, www.cato.org/blog/clean-energy-subsidies-versus-carbon-tax#:~:text=At%20standard%20parameter%20values%2C%20clean,a%20tax%20on%20dirty%20energy. Accessed 3 Sept. 2024.

Moseman, Andrew, and Yogesh Surendranath. "How Can Burning One Ton of Fuel Create More than One Ton of CO2?" *Climate Portal*, MIT, 9 Feb. 2023, climate.mit.edu/ask-mit/how-can-burning-one-ton-fuel-create-more-one-ton-co2#:~:text=For%20example%2C%20says%20Yogesh%20Surendranath,for%20every%20ton%20of%20coal. Accessed 3 Sept. 2024.

Muller, Nicholas Z. "Carbon Tax Effects on Air Quality." *Niskanen Center*, 6 Jan. 2024, www.niskanencenter.org/carbon-tax-effects-on-air-quality/#:~:text=Thus%2C%20metropolitan%20areas%20in%20the,the%20carbon%20tax%20policies%20analyzed. Accessed 28 Aug. 2024.

Muresianu, Alex. "Carbon Taxes in Theory and Practice." *Tax Foundation*, 2 May 2023, taxfoundation.org/research/all/global/carbon-taxes-in-practice/#:~:text=car%20adoption%20elsewhere.-,Macroeconomic%20Effects,use%20of%20the%20tax%20revenue. Accessed 27 Aug. 2024.

Muth, Daniel. "Pathways to Stringent Carbon Pricing: Configurations of Political Economy Conditions and Revenue Recycling Strategies. a Comparison of Thirty National Level Policies." *Ecological Economics*, vol. 214, Dec. 2023, p. 107995, https://doi.org/10.1016/j.ecolecon.2023.107995. Accessed 5 Aug. 2024.

Naef, Alain. "The Impossible Love of Fossil Fuel Companies for Carbon Taxes." *Ecological Economics*, vol. 217, Mar. 2024, p. 108045, https://doi.org/10.1016/j.ecolecon.2023.108045. Accessed 18 Aug. 2024.

"National Energy Balance." *Sustainable Energy Authority of Ireland*, 1 May 2024, www.seai.ie/data-and-insights/seai-statistics/key-publications/national-energy-balance/. Accessed 24 Aug. 2024.

Nuccitelli, Dana. "How Will Carbon Pricing Impact Inflation?" *Citizens' Climate Lobby*, 20 Nov. 2021, citizensclimatelobby.org/blog/policy/how-will-carbon-pricing-impact-inflation/. Accessed 27 Aug. 2024.

Office of Legacy Management. "Environmental Justice History." *Energy.gov*, www.energy.gov/lm/environmental-justice-history. Accessed 30 Aug. 2024.

"OJK International Research Forum, Save the Planet: The Role of Financial Sector to Support Carbon Reduction and Electric Vehicles Development." *OJK Institute*, 25 Sept. 2023, www.ojk.go.id/ojk-institute/en/capacitybuilding/upcoming/3838/ojk-international-research-forum-save-the-planet-the-role-of-financial-sector-to-support-carbon-reduction-and-electric-vehicles-development. Accessed 19 Aug. 2024.

Parliamentary Budget Office. *Carbon Tax Series Part 1 of 3: What Is the Carbon Tax?* Ireland Parliamentary Budget Office, 2024, data.oireachtas.ie/ie/oireachtas/parliamentaryBudgetOffice/2024/2024-02-29_carbon-tax-series-part-1-of-3-what-is-the-carbon-tax_en.pdf. Accessed 24 Aug. 2024.

"Partners." *Climate Leadership Council*, clcouncil.org/about/partners/. Accessed 3 Sept. 2024.

Partnership for Market Readiness. *Carbon Tax Guide: A Handbook for Policy Makers*. Washington, DC, World Bank, 2017, openknowledge.worldbank.org/server/api/core/bitstreams/3f3c5326-7c41-513a-a598-6e8e535e71b9/content. Accessed 20 Aug. 2024.

Pastor, Manuel, et al. *Up in the Air: Revisiting Equity Dimensions of California's Cap-and-Trade System*. USC Equity Research Institute, Feb. 2022,

dornsife.usc.edu/eri/wp-content/uploads/sites/41/2023/01/CAP_and_TRADE_Updated_2020_v02152022_FINAL.pdf. Accessed 26 Aug. 2024.

Patterson, Jacqueline, et al. *Nuts, Bolts, and Pitfalls of Carbon Pricing: An Equity-Based Primer on Paying to Pollute*. National Association for the Advancement of Colored People, July 2021, naacp.org/resources/nuts-bolts-and-pitfalls-carbon-pricing-equity-based-primer-paying-pollute. Accessed 30 Aug. 2024.

Peters, Grayson. "The Social Cost of the Social Cost of Carbon." *Ecology Law Quarterly*, vol. 50, no. 2, 16 Mar. 2024, www.ecologylawquarterly.org/wp-content/uploads/2024/03/50.2-Peters-Note_-internet-Ready.pdf. Accessed 30 Aug. 2024.

Pomerleau, Shuting. "A Carbon Tax Would Help Reduce the Federal Deficit." *Niskanen Center*, 16 Mar. 2023, www.niskanencenter.org/a-carbon-tax-would-help-reduce-the-federal-deficit/#:~:text=A%20carbon%20tax%20would%20raise,billion%20between%202023%20and%202032. Accessed 27 Aug. 2024.

Prest, Brian C., et al. "Social Cost of Carbon Explorer." *Resources for the Future*, 1 Sept. 2022, www.rff.org/publications/data-tools/scc-explorer/. Accessed 14 June 2024.

"Price Greenhouse Gas Emissions and Other Environmental Harms." *Systems Change Lab*, 2022, systemschangelab.org/finance/price-greenhouse-gas-emissions-and-other-environmental-harms#targets. Accessed 29 July 2024.

"Program Review." *The Regional Greenhouse Gas Initiative*, www.rggi.org/program-overview-and-design/program-review. Accessed 26 Aug. 2024.

"Provincial and Territorial Energy Profiles – British Columbia." *Canada Energy Regulator*, www.cer-rec.gc.ca/en/data-analysis/energy-markets/provincial-territorial-energy-profiles/provincial-territorial-energy-profiles-british-columbia.html. Accessed 27 Aug. 2024.

Pulkkinen, Levi. "Washington Climate Activists Disagree about How to Cut Carbon." *Investigate West*, 12 Mar. 2021, www.invw.org/2021/03/12/washington-climate-activists-disagree-about-how-to-cut-carbon/. Accessed 18 Aug. 2024.

"Putting a Price on Carbon." *CDP*, Apr. 2021, www.cdp.net/en/research/global-reports/putting-a-price-on-carbon. Accessed 4 Sept. 2024.

"Q&A: The Social Cost of Carbon." *Carbon Brief*, 14 Feb. 2017, www.carbonbrief.org/qa-social-cost-carbon/. Accessed 9 June 2024.

Recinos, Ada. "Top 6 U.S. Banks Financed Fossil Fuels with $1.8 Trillion since the Paris Agreement; Chase, Citi, and Bank of America Top the List Worldwide." *Sierra Club*, 13 May 2024, www.sierraclub.org/press-releases/2024/05/top-6-us-banks-financed-fossil-fuels-18-trillion-paris-agreement-chase-citi. Accessed 2 Sept. 2024.

"The Regional Greenhouse Gas Initiative Is a Model for the Nation." *NRDC*, 14 July 2021, www.nrdc.org/resources/regional-greenhouse-gas-initiative-model-nation#:~:text=Modeled%20on%20a%20successful%20acid,competitive%20in%20the%20global%20economy. Accessed 26 Aug. 2024.

The Regional Greenhouse Gas Initiative 10 Years in Review. Acadia Center, 2019, acadiacenter.org/wp-content/uploads/2019/09/Acadia-Center_RGGI_10-Years-in-Review_2019-09-17.pdf. Accessed 26 Aug. 2024.

Regulatory Impact Analysis for the Review of the Clean Power Plan: Proposal. U.S. Environmental Protection Agency, Oct. 2017. *EPA*, www.epa.gov/sites/default/files/2017-10/documents/ria_proposed-cpp-repeal_2017-10.pdf. Accessed 15 July 2024.

"Renewables." *En-Roads*, en-roads.climateinteractive.org/scenario.html?v=24.8.0&p16=-0.05&p39=0. Accessed 3 Sept. 2024.

Rennert, Kevin, and Cora Kingdon. "Social Cost of Carbon 101." *Resources for the Future*, 1 Aug. 2019, www.rff.org/publications/explainers/social-cost-carbon-101/. Accessed 8 June 2024.

Reuters. "Chevron Investors Back Proposal for More Emissions Cuts." *Reuters*, 26 May 2021, www.reuters.com/business/energy/chevron-shareholders-approve-proposal-cut-customer-emissions-2021-05-26/?utm_source=newsletter&utm_medium=email&utm_campaign=greenfin&utm_content=2021-06-02. Accessed 20 Aug. 2024.

"RGGI 101 Factsheet." *The Regional Greenhouse Gas Initiative*, Jan. 2024, www.rggi.org/sites/default/files/Uploads/Fact%20Sheets/RGGI_101_Factsheet.pdf. Accessed 26 Aug. 2024.

"The Rise of Sustainable Finance: Green Investing, ESG and Impact on Finance Careers." *Online Business Blog*, Raymond A. Mason School of Business, online.mason.wm.edu/blog/the-rise-of-sustainable-finance#:~:text=Future%20Trends%20in%20Sustainable%20Finance,green%20investing%20practices%20and%20products. Accessed 19 Aug. 2024.

Roy, Nicholas, and Dallas Burtraw. "California's Cap-and-Trade Program and Improvements in Local Air Quality." *Resources*, 4 Oct. 2023, www.resources.org/archives/californias-cap-and-trade-program-and-improvements-in-local-air-quality/. Accessed 26 Aug. 2024.

Running, Katrina. "Towards Climate Justice: How Do the Most Vulnerable Weigh Environment–economy Trade-offs?" *Social Science Research*, vol. 50, Mar. 2015, pp. 217-28. *[Database Name]*, https://doi.org/10.1016/j.ssresearch.2014.11.018.

Sarinsky, Max, and Kurt Weatherford. *The Social Cost of Greenhouse Gases: An Overview: A Primer on EPA's Updated Values for Policymakers and Practitioners*. Institute for Policy Integrity. *JSTOR*, www.jstor.org/stable/resrep60018. Accessed 1 May 2024.

Schimel, Kate. "What Killed Washington's Carbon Tax?" *HighCountry News*, 21 Jan. 2019, www.hcn.org/issues/51-1/energy-and-industry-what-killed-washingtons-carbon-tax/. Accessed 27 Aug. 2024.

Shabecoff, Philip. "Reagan Order on Cost-benefit Analysis Stirs Economic and Political Debate." *New York Times*, 7 Nov. 1981, www.nytimes.com/1981/11/07/us/reagan-order-on-cost-benefit-analysis-stirs-economic-and-political-debate.html. Accessed 9 June 2024.

Sirna, Tony. "Clean Electricity Standards." *Advance Climate Policy*, Citizens' Climate Lobby, 17 Feb. 2021, citizensclimatelobby.org/blog/policy/clean-electricity-standards-vs-carbon-taxes/. Accessed 3 Sept. 2024.

Smith, Adam B. "2023: A Historic Year of U.S. Billion-dollar Weather and Climate Disasters." *Beyond the Data*, Climate.gov, 8 Jan. 2024, www.climate.gov/news-features/blogs/beyond-data/2023-historic-year-us-billion-dollar-weather-and-climate-disasters#:~:text=Adding%20the%202023%20events%20to,376%20events%20exceeds%20%242.660%20trillion. Accessed 2 Sept. 2024.

Smith, E. Keith, et al. "Polarisation of Climate and Environmental Attitudes in the United States, 1973-2022." *Npj Climate Action*, vol. 3, no. 1, 10 Jan. 2024, https://doi.org/10.1038/s44168-023-00074-1. Accessed 2 Sept. 2024.

Smith, Tori. "U.S. Carbon Border Adjustment Proposals and World Trade Organization Compliance." *American Action Forum*, 8 Feb. 2023, www.americanactionforum.org/insight/u-s-carbon-border-adjustment-proposals-and-world-trade-organization-compliance/. Accessed 22 Aug. 2024.

"States Using the SC-GHG." *The Cost of Climate Pollution*, costofcarbon.org/states. Accessed 3 Sept. 2024.

Stavins, Robert N. "The Future of US Carbon-Pricing Policy." *Environmental and Energy Policy and the Economy*, vol. 1, Jan. 2020, pp. 8-64, https://doi.org/10.1086/706792. Accessed 2 Sept. 2024.

Sterner, Thomas, et al. "Economists and the Climate." *Journal of Behavioral and Experimental Economics*, vol. 109, Apr. 2024, p. 102158, https://doi.org/10.1016/j.socec.2023.102158. Accessed 18 Aug. 2024.

Stock, James H., and Gilbert E. Metcalf. *The Macroeconomic Impact of Europe's Carbon Taxes*. Environment for Development Initiative, 1 Jan. 2020. *JSTOR*, www.jstor.org/stable/resrep46948. Accessed 27 Aug. 2024.

Stuart, Daniel, and Paul Hibbard. *The Economic Impacts of the Regional Greenhouse Gas Initiative on Ten Northeast and Mid-Atlantic States*. Analysis Group, May 2023, www.analysisgroup.com/globalassets/insights/publishing/2023-ag-rggi-report.pdf. Accessed 26 Aug. 2024.

Swanson, Conrad. "Federal Judge Dismisses Lawsuit against WA's Carbon-Pricing Law." *The Seattle Times*, 14 Nov. 2023,

www.seattletimes.com/seattle-news/environment/federal-judge-dismisses-lawsuit-against-was-carbon-pricing-law/. Accessed 27 Aug. 2024.

---. "Initiative 2117 to Repeal WA Climate Act Takes Key Step toward Ballot." *The Seattle Times*, 16 Jan. 2024, www.seattletimes.com/seattle-news/environment/initiative-2117-to-repeal-wa-climate-act-takes-key-step-toward-ballot/. Accessed 27 Aug. 2024.

Temple, James, and Lisa Songarchive. "The Climate Solution Actually Adding Millions of Tons of CO2 into the Atmosphere." *MIT Technology Review*, 29 Apr. 2021, www.technologyreview.com/2021/04/29/1017811/california-climate-policy-carbon-credits-cause-co2-pollution/. Accessed 28 Aug. 2024.

"Top 10 Climate Tech Trends and Innovations in 2025." *StartUs Insights*, Aug. 2024, www.startus-insights.com/innovators-guide/climate-tech-trends-innovations/. Accessed 4 Sept. 2024.

"Total U.S. Greenhouse Gas Emissions by Economic Sector in 2022." *EPA*, www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions. Accessed 28 Aug. 2024. Map.

"Trump Vs. Obama on the Social Cost of Carbon–and Why It Matters." *Center on Global Energy Policy*, 15 Nov. 2017, www.energypolicy.columbia.edu/publications/trump-vs-obama-social-cost-carbon-and-why-it-matters/. Accessed 3 Sept. 2024.

"2020 Household Impact Study." *Citizens' Climate Lobby*, citizensclimatelobby.org/household-impact-study/?_ga=2.238774403.220757472.1721316787-408811361.1721316787. Accessed 28 Aug. 2024.

"Two Years Later: The Impact of the Inflation Reduction Act on Environmental Justice." *North Carolina Black Alliance*, 8 Aug. 2024, ncblackalliance.org/two-years-later-the-impact-of-the-inflation-reduction-act-on-environmental-justice/#:~:text=Why%20is%20IRA%20important%20for,%2C%20pollution%20reduction%2C%20and%20weatherization. Accessed 30 Aug. 2024.

Tyson, Alec, and Brian Kennedy. "Two-Thirds of Americans Think Government Should Do More on Climate." *Pew Research Center*, 23 June 2020, www.pewresearch.org/science/2020/06/23/two-thirds-of-americans-think-government-should-do-more-on-climate/. Accessed 2 Sept. 2024.

Ummel, Kevin. *2020 Household Impact Study*. Citizens' Climate Lobby, Aug. 2020, citizensclimatelobby.org/wp-content/uploads/2018/06/HIS2-Working-Paper-v1.1.pdf. Accessed 28 Aug. 2024.

United States, U.S. Supreme Court (U.S.). *Massachusetts v. EPA*. *United States Reports*, vol. 549, 2 Apr. 2007. *Casetext*, Thomson Reuters, casetext.com/case/massachusetts-v-environmental-protection-agency-1. Accessed 15 July 2024.

United States, Congress, House, Committee on Oversight and Reform. *Analysis of the Fossil Fuel Industry's Legislative Lobbying and Capital Expenditures Related to Climate Change*. 28 Oct. 2021, oversightdemocrats.house.gov/sites/evo-subsites/democrats-oversight.house.gov/files/Analysis%20of%20the%20Fossil%20Fuel%20Industrys%20Legislative%20Lobbying%20and%20Capital%20Expenditures%20Related%20to%20Climate%20Change%20-%20Staff%20Memo%20%2810.28.21%29.pdf. Accessed 18 Aug. 2024. 117th Congress.

---, ---, House. Energy Innovation and Carbon Dividend Act of 2018. *Congress.gov*, www.congress.gov/bill/115th-congress/house-bill/7173#:~:text=Energy%20Innovation%20and%20Carbon%20Dividend%20Act%20of%202018,-This%20bill%20amends&text=The%20bill%20also%20imposes%20a,and%20sequester%20carbon%20dioxide%2C%20and. Accessed 2 Sept. 2024. 115th Congress, House Bill 7173.

---, ---, House, Committee on Oversight and Reform. *The Devastating Impacts of Climate Change on Health*. Testimony of Drew Shindell, Distinguished Professor of Earth Sciences, Nicholas School of the Environment, Duke University. *Govinfo.gov*, 5 Aug. 2020, www.govinfo.gov/content/pkg/CHRG-116hhrg41911/html/CHRG-116hhrg41911.htm. Accessed 2 Sept. 2024. 115th Congress, 2nd session.

---, ---, Senate, Committee on the Budget. *Who Pays the Price: The Real Cost of Fossil Fuels*. Testimony of Sheldon Whitehouse, Chairman, U.S. Senate Budget Committee. 3 May 2023, www.budget.senate.gov/chairman/newsroom/press/sen-whitehouse-on-fossil-fuel-subsidies-we-are-subsidizing-the-danger-#:~:text=It%27s%20not%20just%20the%20US,to%20the%20fossil%20fuel%20industry. Accessed 2 Sept. 2024.

---, ---, Senate. Foreign Pollution Fee Act of 2023. *Congress.gov*, www.congress.gov/bill/118th-congress/senate-bill/3198/text#:~:text=Introduced%20in%20Senate%20(11%2F02%2F2023)&text=To%20amend%20the%20Internal%20Revenue,products%2C%20and%20for%20other%20purposes. Accessed 27 Aug. 2024. 118th Congress, 1st session, Senate Bill 3198, referred to Senate Committee 2 Nov. 2023.

---, Executive Office of the President [Donald J. Trump]. Executive Order 13783: Promoting Energy Independence and Economic Growth. 28 Mar. 2017. *Federal Register*, vol. 82, 31 Mar. 2017, pp. 16093-97, www.federalregister.gov/documents/2017/03/31/2017-06576/promoting-energy-independence-and-economic-growth. Accessed 15 July 2024.

"United States." *IEA*, www.iea.org/reports/world-energy-investment-2024/united-states. Accessed 2 Sept. 2024.

United States, Ninth Circuit Court (9th Cir.). *Center for Biological Diversity v. National Highway Traffic Safety Administration*. *Federal Reporter, Third Series*, vol. 538, 18 Aug. 2008, casetext.com/case/center-for-biological-v-nhtsa/. Accessed 15 July 2024.

United States Climate Alliance. usclimatealliance.org/about/. Accessed 20 Aug. 2024.

"USA - Washington Cap-and-invest Program." *International Carbon Action Partnership*, icapcarbonaction.com/en/ets/usa-washington-cap-and-invest-program. Accessed 26 Aug. 2024.

"U.S. Energy Facts Explained." *U.S. Energy Information Administration*, www.eia.gov/energyexplained/us-energy-facts/#:~:text=U.S.%20total%20annual%20energy%20production,Click%20to%20enlarge. Accessed 24 Aug. 2024.

"Virginia Fails to Rejoin RGGI." *Environment Virginia*, 15 May 2024, environmentamerica.org/virginia/updates/virginia-fails-to-rejoin-rggi/. Accessed 26 Aug. 2024.

Warner, Mandy. "With the Support of a Majority of Pennsylvanians, State Officially Links to the Regional Greenhouse Gas Initiative." *Environmental Defense Fund*, 22 Apr. 2022, www.edf.org/media/support-majority-pennsylvanians-state-officially-links-regional-greenhouse-gas-initiative. Accessed 26 Aug. 2024.

"Washington's Cap-and-Invest Program." *Department of Ecology State of Washington*, ecology.wa.gov/air-climate/climate-commitment-act/cap-and-invest. Accessed 26 Aug. 2024.

Weiss, Daniel J. "Anatomy of a Senate Climate Bill Death." *Center for American Progress*, 12 Oct. 2010, www.americanprogress.org/article/anatomy-of-a-senate-climate-bill-death/. Accessed 2 Sept. 2024.

"What Is a Carbon Tax? How Would It Affect the Economy?" *Peter G. Peterson Foundation*, 30 Sept. 2021, www.pgpf.org/budget-basics/what-is-a-carbon-tax-how-would-it-affect-the-economy. Accessed 27 Aug. 2024.

Wherry, Aaron. "Is the Carbon Tax Suffering from a Failure to Communicate?" *CBC*, 14 Mar. 2024, www.cbc.ca/news/politics/carbon-tax-trudeau-premiers-1.7142878. Accessed 24 Aug. 2024.

The White House. "Fact Sheet: Biden-Harris Administration Announces New Actions to Reduce Greenhouse Gas Emissions and Combat the Climate Crisis." 21 Sept. 2023, www.whitehouse.gov/briefing-room/statements-releases/2023/09/21/fact-sheet-biden-harris-administration-announces-new-actions-to-reduce-greenhouse-gas-emissions-and-combat-the-climate-crisis/. Accessed 18 July 2024.

---. "Fact Sheet: President Biden Sets 2030 Greenhouse Gas Pollution Reduction Target Aimed at Creating Good-Paying Union Jobs and Securing U.S. Leadership on Clean Energy Technologies." 22 Apr. 2021, www.whitehouse.gov/briefing-room/statements-releases/2021/04/22/fact-sheet-president-biden-sets-2030-greenhouse-gas-pollution-reduction-target-aimed-at-creating-good-paying-union-jobs-and-securing-u-s-leadership-on-clean-energy-technologies/#:~:text=On%20Day%20One%2C%20President%20Biden,by%20no%20later%20than%202050. Accessed 2 Sept. 2024.

"Who Supports a Price on Carbon?" *Citizens' Climate Lobby*, citizensclimatelobby.org/who-supports-a-price-on-carbon/. Accessed 18 Aug. 2024.

"William D. Nordhaus Biographical." *The Nobel Prize*, www.nobelprize.org/prizes/economic-sciences/2018/nordhaus/biographical/. Accessed 3 Sept. 2024.

Wolman, Jordan, et al. "Republicans Are Trying to Snuff out Climate Embers around the Country." *Politico*, 24 Apr. 2024, www.politico.com/news/2024/04/24/democrat-climate-policies-populist-backlash-00153601. Accessed 24 Aug. 2024.

World Bank. *State and Trends of Carbon Pricing 2024*. Washington D.C., Open Knowledge Repository, 21 May 2024,

openknowledge.worldbank.org/server/api/core/bitstreams/253e6cdd-9631-4db2-8cc5-1d013956de15/content. Accessed 28 July 2024.

World Bank Group. "Prices in ETSs and Carbon Taxes in 2024." World Bank, carbonpricingdashboard.worldbank.org/compliance/revenue. Accessed 29 July 2024. Chart.

"World Energy Investment 2023." *International Energy Agency*, 2023, www.iea.org/data-and-statistics/charts/increase-in-annual-clean-energy-investment-in-selected-countries-and-regions-2019-2023. Accessed 22 Aug. 2024.

"WTO Chief Calls for Global Carbon Price, Reforms to Tariffs and Red Tape to Clean up Supply Chains." *World Economic Forum*, 20 Jan. 2023, www.weforum.org/press/2023/01/wto-chief-calls-for-global-carbon-price-reforms-to-tariffs-and-red-tape-to-clean-up-supply-chains/#:~:text=The%20WTO%20is%20working%20with,%2C%E2%80%9D%20said%20Okonjo%2DIweala. Accessed 22 Aug. 2024.

Yale University, and Second Nature, editors. "Internal Carbon Pricing in Higher Education Toolkit." *Second Nature*, secondnature.org/resources/offsets/internal-carbon-pricing-in-higher-education-toolkit/. Accessed 4 Sept. 2024.

Ye, Jason. "Carbon Pricing Proposals in the 117th Congress." *Center for Climate and Energy Solutions*, Dec. 2022, www.c2es.org/wp-content/uploads/2021/12/carbon-pricing-proposals-in-the-117th-congress.pdf. Accessed 2 Sept. 2024.

---. *Options and Considerations for a Federal Carbon Tax*. Center for Climate and Energy Solutions, Feb. 2013, www.c2es.org/wp-content/uploads/2013/02/options-considerations-federal-carbon-tax.pdf. Accessed 20 Aug. 2024.

Yin, Hua-Tang, et al. "Carbon Tax: Catalyst or Hindrance for Renewable Energy Use in Climate Change Mitigation?" *Energy Strategy Reviews*, vol. 51, Jan. 2024, p. 101273, https://doi.org/10.1016/j.esr.2023.101273. Accessed 22 Aug. 2024.

Yoder, Kate. "Washington's Key Climate Law Is under Attack. Big Oil Wants It to Survive." *Grist*, 13 Feb. 2024, grist.org/politics/washington-cap-and-invest-law-repeal-oil-companies/. Accessed 27 Aug. 2024.

Zachmann, Georg, and Ben McWilliams. *A European Carbon Border Tax: Much Pain, Little Gain*. 1 Mar. 2020, www.jstor.org/stable/resrep28625. Accessed 27 Aug. 2024.

Zakrzewski, Katie. "Carbon Border Adjustment Mechanisms: What Are They, and Why Do They Matter?" *Citizens' Climate Lobby*, 1 Sept. 2022, citizensclimatelobby.org/blog/policy/carbon-border-adjustment-mechanisms-what-are-they-and-why-do-they-matter/. Accessed 27 Aug. 2024.

---. "Carbon Fee vs Carbon Tax: What's the Difference?" *Advance Climate Policy*, Citizens' Climate Lobby, 7 June 2022, citizensclimatelobby.org/blog/policy/carbon-fee-vs-carbon-tax/. Accessed 2 Sept. 2024.

---. "A Carbon Price Is Better than Carbon Credits." *Citizens' Climate Lobby*, 30 June 2022, citizensclimatelobby.org/blog/policy/a-carbon-price-is-better-than-carbon-credits/. Accessed 28 Aug. 2024.

Predicting Mental Health and Mood Swings Based on Demographic, Lifestyle, and Emotional Factors Using Deep Learning and Neural Networks By Aaryan Sharma

Abstract

The COVID-19 pandemic has exacerbated mental health challenges, with anxiety and depression surging globally. This study explores mood swings as an early indicator of mental health issues, utilizing a deep neural network to predict mood variability. The independent variables (predictors) we used to identify mood swings and mental health issues include demographic data such as age, gender, and employment status, which are known to influence mental health lifestyle factors such as daily habits, work-life balance, and social support levels and emotional states such as self-reported stress, anxiety, and depression levels. The Kaggle Mental Health Dataset was preprocessed by handling missing values, removing duplicates, encoding categorical features, and standardizing variables.

A ReLU-based neural network model was developed, utilizing demographic, lifestyle, and emotional state variables to uncover complex, non-linear relationships affecting mood swings. The study demonstrates strong predictive performance, achieving an accuracy of 83%. This research underscores the potential of deep learning and neural networks to enhance early diagnosis and personalized mental health interventions. Quantitative results and trends support the model's robustness, offering a data-driven approach to addressing mental health crises.

Introduction

In today's world, while remarkable advancements in science and technology mark human progress, pressing issues like mental health remain a growing concern. The COVID-19 pandemic has significantly worsened this situation. For instance, the World Health Organization (WHO) reported a 25% increase in the global prevalence of anxiety and depression during the pandemic's first year ("COVID-19 Pandemic Triggers 25% Spike"). This surge can be attributed to social isolation, economic insecurity, and healthcare disruptions (Panchal et al.). This stark reality underscores the urgent need to prioritize mental health awareness, access to care, and societal resilience.

Mood swings, a common and significant feature of various psychiatric disorders, serve as an early warning sign for potential mental health problems ("Adult Psychiatric Morbidity Survey"). Studies suggest that mood instability, particularly pronounced during adolescence, is associated with a variety of mental health conditions, including depression, anxiety, and bipolar disorder, and is linked to increased health service utilization and suicidal ideation (Marwaha et al.). Mood instability is reported in 40–60% of those with depression, anxiety disorder, post-traumatic stress disorder, and obsessive-compulsive disorder. It is associated with increased health service use and suicidal ideation, independent of neurotic symptoms, alcohol misuse, borderline personality disorder, and other confounders.

Building on this understanding, this paper leverages advancements in deep learning and neural networks to predict and quantify mood swings as a step toward addressing broader mental

health concerns. Utilizing self-reported patient data, machine learning models are designed to identify predictive factors for mood instability, which has transdiagnostic potential as both an investigational and therapeutic target. By analyzing neurobiological correlates, prevalence data, and clinical characteristics, these models aim to improve early diagnosis and personalized intervention strategies. Such tools not only enhance our ability to monitor mental health conditions but also hold promise in mitigating the long-term impacts of psychiatric disorders, including bipolar disorder, borderline personality disorder, and psychotic disorders. Integrating AI technology into mental health research exemplifies how data-driven approaches can bridge the gap between clinical understanding and proactive care. By focusing on mood instability as a key metric, we can advance therapeutic outcomes and contribute to a deeper understanding of mental health dynamics.

Materials and Methods

In this study, we utilized the Mental Health Dataset from Kaggle, which provides comprehensive data on mental health indicators, including self-reported emotional states, demographic information, and related lifestyle factors. This dataset offers a valuable foundation for exploring mental health patterns and predictive modeling, particularly in identifying mood swings. The Kaggle Mental Health Dataset contains survey responses on mental health issues, including emotional states, lifestyle habits, and personal well-being factors. Key insights from the dataset show that a significant portion of respondents report experiencing stress, anxiety, and depression. Additionally, factors such as age, employment status, and social support influence mental health outcomes. This data is ideal for machine learning models that predict mood swings and identify mental health patterns across diverse demographic groups. Below is a detailed description of the independent variables that predict mood swings, including their data characteristics. These variables were preprocessed (e.g., standardized or encoded) to enhance compatibility with the neural network and ensure robust predictions.

1. **Age**:
    - **Data Type**: Numerical, continuous.
    - **Example Values**: 18, 25, 45, 60.
    - **Description**: Captures the respondent's age. Young adults and older populations often show differing patterns of mood instability.
2. **Gender**:
    - **Data Type**: Categorical, encoded numerically (e.g., 0 = Male, 1 = Female, 2 = Other).
    - **Example Values**: 0, 1, 2.
    - **Description**: Identifies gender-based trends in mental health.
3. **Employment Status**:
    - **Data Type**: Categorical, encoded numerically (e.g., 0 = Unemployed, 1 = Employed, 2 = Self-employed).

    - **Example Values**: 0, 1, 2.
    - **Description**: Assesses how professional stability or lack thereof impacts mood.
4. **Emotional States (Stress, Anxiety, Depression)**:
    - **Data Type**: Numerical, scaled from self-reported surveys (e.g., 0–10).
    - **Example Values**: Stress = 7, Anxiety = 5, Depression = 8.
    - **Description**: Quantifies individual mental health conditions directly tied to mood variability.
5. **Lifestyle Factors**:
    - **Data Type**: Numerical, based on surveys or encoded metrics (e.g., hours of sleep, physical activity levels).
    - **Example Values**: Sleep = 6 hours, Physical Activity = 3 days/week.
    - **Description**: Measures behaviors influencing mental health.
6. **Occupation**:
    - **Data Type**: Categorical, encoded numerically (e.g., 0 = IT, 1 = Healthcare, 2 = Retail).
    - **Example Values**: 0, 1, 2.
    - **Description**: Accounts for work-related stress and its impact.
7. **Relationship Status**:
    - **Data Type**: Categorical, encoded numerically (e.g., 0 = Single, 1 = Married, 2 = Divorced).
    - **Example Values**: 0, 1, 2.
    - **Description**: Represents social support or isolation factors.

To analyze this data, we implemented a ReLU-based neural network model to predict mood swings, focusing on uncovering the complex, non-linear relationships between input features and mood instability. Mood swings were measured quantitatively using self-reported survey data from the Mental Health Dataset. Participants rated their emotional states on a categorical numeric scale ranging from 0(low mood instability) , 1 (moderate mood instability) ,and 2 (high mood instability) . These scores were derived from responses to questions about stress, anxiety, depression, and other mood-related factors. Additionally, lifestyle and demographic data were analyzed to correlate with these ratings, enabling the identification of patterns and predictors of mood variability. The scores served as input for the neural network model. Mood swings, a significant marker for mental health conditions, were identified as a key target for this predictive approach. By training the neural network on this dataset, we aimed to enhance the understanding of mood variability and its potential as an early indicator of mental health challenges.

This research bridges deep learning methodologies with mental health analytics, demonstrating the potential for data-driven approaches to improve diagnostic precision and support mental health interventions.

Neural Network Architecture

Neural networks mimic how the human brain processes information to recognize patterns and make predictions. It consists of layers of interconnected nodes (neurons), starting with the input layer, which takes raw data (like pixels of an image). The data is passed through hidden layers, where each neuron uses mathematical functions to focus on specific features, such as edges or shapes. ReLU (Rectified Linear Unit) is commonly used here to make the network capable of capturing complex patterns by activating only relevant signals and ignoring the rest. Finally, the output layer provides the prediction, and feedback helps the network adjust its internal parameters to improve accuracy over time.

More specifically, a neural network works like a digital brain that learns through trial and error. It starts with random "weights" to determine how strong the connections between its "neurons" should be. In hidden layers, weighted sums of inputs are calculated, and activation functions like ReLU highlight patterns by introducing non-linearity. Input data flows through these neurons in layers, where calculations combine the inputs, apply functions like ReLU to highlight patterns, and generate a prediction. Data flows through input, hidden, and output layers and then compares this prediction to the actual value using a loss function to measure errors. Adjustments (backpropagation) refine the weights repeatedly through training cycles until the network learns to make accurate predictions. Training iterates over epochs, with data divided into batches for efficiency, refining predictions until the loss stabilizes.

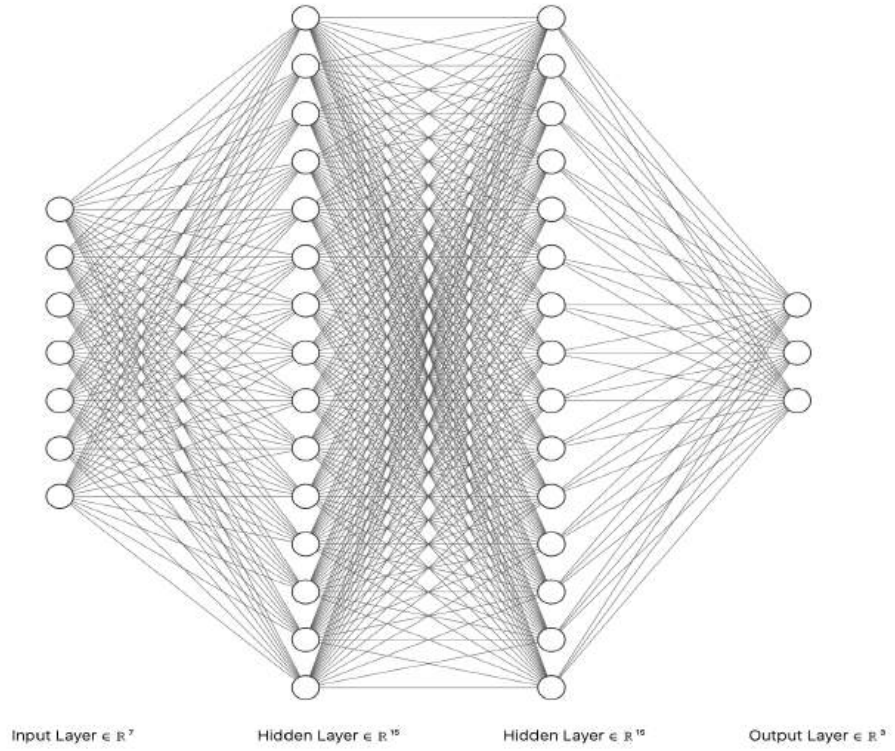


Figure 1: Design of a Neural Network

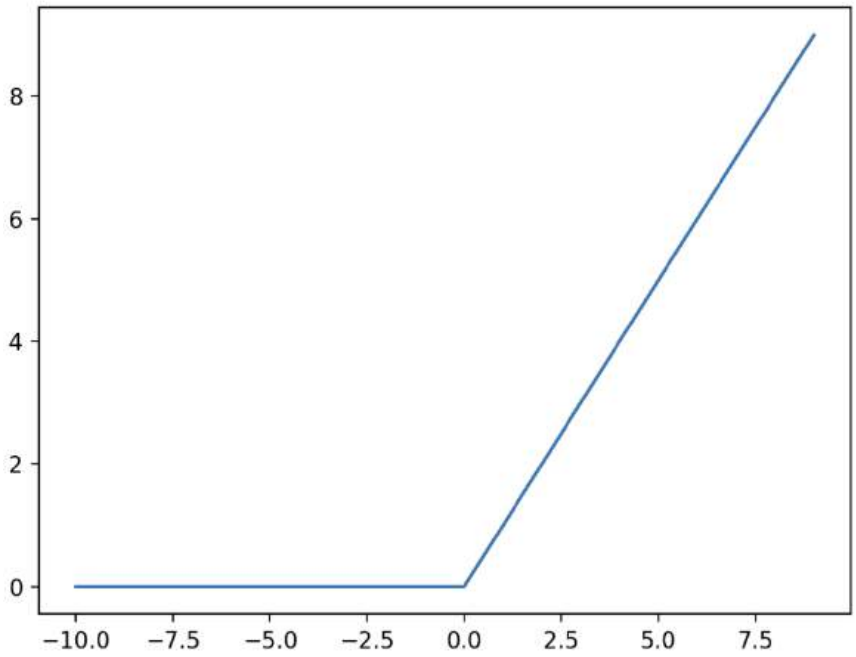


Figure 2: ReLU Activation Function

Below is a detailed summary of steps describing the mathematical intuition behind neural networks:

1. **Weight Initialization**: At the start, all connections (edges) between neurons are assigned small, random weights. These weights determine how much influence one neuron has on the next. Random initialization ensures that neurons learn different features.
2. **Feedforward Propagation**:
    - **Input Layer**: The raw input data is introduced to the network. Each feature is represented as a numerical value and fed into the input nodes.
    - **Hidden Layers**: Data flows sequentially through hidden layers. At each neuron, a weighted sum of the inputs is computed:

$$z = \sum w_i x_i + b$$

      where w is the weight, x is the input, and b is the bias term.
    - **Activation Functions**: Non-linear activation functions, like $ReLU\ max(0,\ z)$, are applied to the weighted sum, introducing non-linearity. This enables the network to learn complex patterns.
    - **Output Layer**: The final layer produces the network's predictions (y_{pred}) which might represent probabilities (for classification) or continuous values (for regression).

3. **Loss Function**: The loss function measures the difference between the network's prediction (y_{pred}) and the actual output (y_i). For instance, Mean Squared Error is common in regression tasks, while cross-entropy loss is typical in classification tasks. The goal is to minimize this loss.
4. **Backward Propagation (Backprop)**:
    - **Gradient Computation**: Gradients (partial derivatives) of the loss with respect to weights and biases are calculated using the chain rule. This determines how each weight contributes to the error.
    - **Weight Updates**: Using an optimization algorithm (e.g., stochastic gradient descent or Adam), weights and biases are updated to minimize the loss:

      $w_{new} = w_{old} - \eta . \frac{\partial L}{\partial w}$

      where , η is the learning rate, which controls the step size for the update, and $\frac{\partial L}{\partial w}$ is the gradient of the loss L with respect to the weight w is the learning rate.
5. **Training Iterations**:
    - Feedforward propagation, loss calculation, and backpropagation are repeated over multiple iterations (epochs). During each epoch, the entire dataset is passed through the network.
    - Data is divided into smaller subsets (batches) to manage computational efficiency. Each batch updates the model weights, ensuring faster convergence and less memory usage.
6. **Convergence**: Training continues until the loss stabilizes or validation metrics (e.g., accuracy or validation loss) stop improving, indicating that the model has learned as much as possible from the data.

Our Methodology

Neural networks are computational models inspired by the structure of the human brain, consisting of layers of interconnected neurons that process and learn from input data. Each neuron performs a mathematical operation on the input, and the network adjusts its internal parameters through training to minimize prediction errors. **ReLU (Rectified Linear Unit)** is a commonly used activation function in neural networks, introducing non-linearity and helping the model capture complex patterns by transforming negative values to zero while passing positive values unchanged. This makes the network capable of learning intricate relationships in data. As part of our research, we leveraged a **ReLU neural network** model for predicting mood swings based on a mental health dataset. We followed the steps listed below:

1. Data Preprocessing

In the data preprocessing step, missing values in the 'self_employed' column are addressed by removing rows with null values. Duplicated rows are identified and removed to maintain the integrity of the dataset. Categorical features are then transformed into numerical

values using **LabelEncoder**, allowing the neural network to process them. Finally, **StandardScaler** is applied to normalize the feature values, ensuring all input variables are on a similar scale, which improves model performance and convergence during training. These preprocessing steps help clean and prepare the data for effective model training.

1. **Handling Missing Values**: The dropna() function is used to remove rows containing missing values. Specifically, rows with null values in the 'self_employed' column are dropped to ensure the dataset remains complete and valid for analysis.
2. **Removing Duplicates**: The duplicated() function identifies duplicate rows, and drop_duplicates() removes these rows to prevent redundant data that could skew model performance.
3. **Label Encoding**: The LabelEncoder() function converts categorical variables (e.g., 'Gender', 'Occupation') into the numerical format by assigning each category a unique integer. This step is essential for neural networks, which require numeric input.
4. **Standard Scaling**: The StandardScaler() standardizes the features by removing the mean and scaling them to unit variance, ensuring that all features contribute equally during training. This prevents the model from being biased towards variables with larger ranges.

2. Model Development

We built a **sequential neural network with three hidden layers**. A sequential neural network is designed by stacking multiple layers of neurons, each performing a computation on the input data. In this model, three hidden layers with 50 neurons each allow the network to learn increasingly complex data representations. This architecture enables the model to capture deep patterns in the data that may not be obvious with simpler structures.

The **input features** in our dataset are: **Gender, Occupation, Family History, Days Indoors, Changes in Habits, Mental Health History,** and **Social Weakness.** The main target variable is Mood Swings. Therefore, the model consists of **7 input features**, **3 hidden layers (each with 50 neurons)**, and **1 output neuron**.

a. **Input Layer:** The neural network starts with the input layer, which takes in the 7 features. This forms a vector of size 7, representing the input to the first layer of the network.

$$X = \left[x_1, x_2, x_3, x_4, x_5, x_6, x_7\right]$$

Where:

- x_1=Gender
- x_2=Occupation
- x_3=Family History

- x_4=Days Indoors
- x_5=Changes in Habits
- x_6=Mental Health History
- x_7=Social Weakness

b. **Hidden Layers:** The network contains **three hidden layers**. **The ReLU** (Rectified Linear Unit) activation function (Figure 2 above) is used in the hidden layers to introduce non-linearity. This helps the model learn complex, non-linear relationships between the input features and the output, making it more powerful in capturing patterns that simpler linear models might miss. The first hidden layer is defined as below. It has 50 neurons and uses the ReLU activation function.

"layers.Dense(units=50, activation='relu', input_shape=[7])"

Each neuron in the first hidden layer calculates a weighted sum of the inputs, adds a bias, and applies the ReLU activation function. For the j-th neuron in the first hidden layer:

$$z_1^{(1)} = \sum_{i=1}^{7} w_{ij}^{(1)}.x_i + b_j^{(1)}$$

$$a_j^{(1)} = ReLU\left(z_j^{(1)}\right) = max\left(0, z_j^{(1)}\right)$$

Where:

- $w_{ij}^{(1)}$ is the weight from input i to neuron j in the first hidden layer.
- x_i is the input value for the i-th feature.
- $b_j^{(1)}$ is the bias for neuron j in the first hidden layer.
- $z_1^{(1)}$ is the pre-activation value for neuron j in the first hidden layer
- $a_j^{(1)}$ is the activation (output) of neuron j in the first hidden layer

This is repeated for all 50 neurons in the first hidden layer, forming an output vector $a_j^{(1)}$ of size 50. The second and third hidden layers are defined as below. Each has 50 neurons and uses the ReLU activation function

"layers.Dense(units=50, activation='relu')"

For the second and third hidden layers (with 50 neurons each), the process is similar:

$$z_k^{(l)} = \sum_{j=1}^{50} w_{jk}^{(l)} a_j^{(l-1)} + b_k^{(l)}$$

$$a_k^{(l)} = ReLU\left(z_k^{(l)}\right) = max\left(0, z_k^{(l)}\right)$$

Where

- $w_{jk}^{(l)}$ is the weight between the j-th unit in the previous layer and the k-th unit in the current layer.
- $a_j^{(l-1)}$ is the activation from the j-th unit in the previous layer (layer l - 1).
- $b_k^{(l)}$ is the bias term for the k-th unit in layer l.
- The summation runs over all 50 previous units.
- $a_k^{(l)}$ is the activation of the k-th unit in layer l, l is the layer index (l=2,3)
- **ReLU** (Rectified Linear Unit) applies the function $max\left(0,\ z_k^{(l)}\right)$, where values less than 0 are set to 0.

c. **The output layer** (with Softmax activation) calculates a weighted sum of the activations from the last hidden layer:

$$z_k^{(L)} = w_k^{(L)} a_k^{(L-1)} + b_k^{(L)} \text{ for each output unit K = 1, 2 3}$$

Where:

- $w_k^{(L)}$ is the weight connecting neuron j in the third hidden layer to the output neuron. $z_k^{(L)}$ is the sum of weighted inputs $w_k^{(L)}$ and $a_k^{(L-1)}$, with a bias $b_k^{(L)}$

The **Softmax** function is then applied to convert these logits into class probabilities

$$p_k = \frac{e^{z_k^{(L)}}}{\sum_{j=1}^{3} e^{z_j^{(L)}}}$$

- p_k is the probability of class k
- $z_k^{(L)}$ is the raw score for class k

The predicted class is the one with the highest probability:

$$\hat{y}_i = arg\ max\ (p1,\ p2,\ p3)$$

Where $\hat{y}_i$ is the predicted class for the i^{th} sample.

3. Training & Validation

At a high level, in this step, the dataset is split into training and validation sets using **train_test_split**, typically allocating 80-90% of data for training and the remaining 10-20% for validation. **10 epochs** allow the model to iterate over the dataset multiple times, each epoch refining the model's weights. A **batch size of 40** means the model processes 40 samples at once

before updating weights, balancing computational efficiency and effective learning. Monitoring **validation loss** ensures the model generalizes well, preventing overfitting if loss increases on the validation set. Below is the stepwise summary.

a. **Loss function:** The Adam optimizer adjusts the learning rate dynamically during training, improving model convergence and efficiency. This is combined with the Categorical Cross-Entropy (CCE) loss function, which penalizes the model for making significant errors in predictions. The loss function is a crucial component of the training process, as it quantifies how well the model's predictions match the actual labels. In this classification task, we use CCE since the model predicts three different classes.

$$\text{Loss}=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\left(y_{ij}log\ (\hat{y}_{ij})\right)$$

Where:

1. N is the number of samples in a batch.
2. C= Number of classes (3 in this case)
3. y_{ij} is the actual class label.
4. $(\hat{y}_{ij})$ is the predicted value for class j.

b. Optimization: The Adam optimizer updates the weights and biases based on the gradients of the loss function:

$$\theta \leftarrow \theta - \eta \cdot \nabla_{\theta}\text{Loss}$$

Where:

- θ represents the parameters (weights and biases).
- η is the learning rate.
- ∇_{θ}Loss is the gradient of the loss with respect to θ.

The choice of **10 epochs** is based on balancing adequate learning without overfitting. If training continues beyond a point where the validation loss stops improving, it could lead to diminishing returns, suggesting the model has learned as much as it can from the data.

4. Model Evaluation

After training the model, predictions $\hat{y}_i$ on the validation set were compared to the actual values y_i. Key metrics used for evaluation include **accuracy**, **precision**, **recall**, and **F1-score**. These metrics are relevant because we are dealing with a classification problem and using **Categorical Cross-Entropy** as the loss function.

Accuracy is the most important metric in this case, as it measures how well the model predicts the correct class:

$$Accuracy\ =\ \frac{\sum_{i=1}^{N}1\,(\hat{y}_i = y_i)}{N}$$

Where:

- N is the total number of samples,
- $1\ (\hat{y}_i = y_i)$ is an indicator function that equals 1 if the predicted class $\hat{y}_i$ matches the true class y_i, and 0 otherwise.

Precision, Recall, and F1-Score are also computed for each class to understand how the model performs across all classes, especially if the class distribution is imbalanced.
A Loss vs. Validation Loss Plot is generated to visualize the model's learning process. A steady decrease in both training and validation loss is desirable, indicating that the model is effectively learning and generalizing to the validation set.
This methodology, combining robust preprocessing, mathematical rigor, and iterative optimization, allowed us to build a model capable of predicting mood swings from mental health data with high precision.

5. Fine-tuning the model

We also experimented with 14-7-7-3 (Model 1) neuron architecture and 50-50-50-3 (Model 2) neurons architecture to find the model with optimal performance. Overall, Model 2 has higher accuracy (83%) compared to Model 1 with 66% accuracy. The larger architecture (Model 2) allows it to capture more complex patterns, leading to better performance across all metrics. Model 1 is at risk of underfitting due to its smaller architecture, while Model 2 achieves stable, high performance with a good balance of precision, recall, and F1-Score. Model 2 also benefits from higher capacity, which enhances its ability to understand complex relationships, whereas Model 1 remains more basic and struggles to capture complex patterns effectively. Table of Results below.

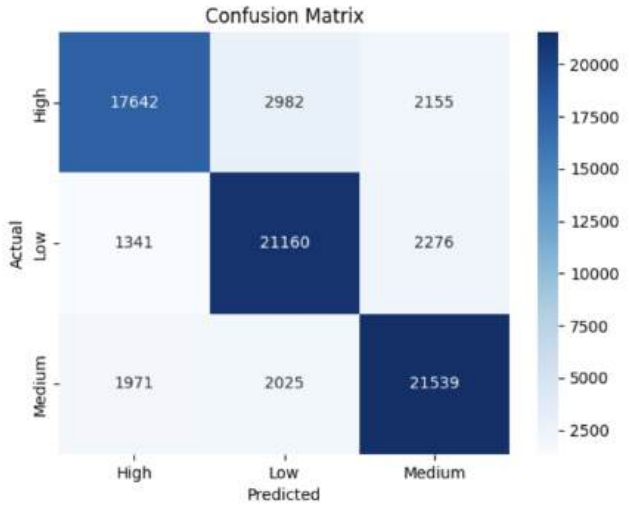
Metric	**Model 1 (14-7-7-3)**	**Model 2 (50-50-50-3)**
Architecture	*(14-7-7-3) 31 neurons*	*(50-50-50-3) 153 neurons*
Accuracy	*66%*	*83%*
Final Training Loss	*0.7191*	*0.3033*
Minimum Validation Loss	*0.7031*	*0.2983*
Validation loss trend	*Fluctuates after epoch 8*	*Steady decline, minor fluctuations*
Risk of Overfitting	*Low (underfitting risk due to lower complexity)*	*Possible (overfitting risk due to model complexity)*
Model Capacity	*Lower capacity, might miss some complex patterns*	*Higher capacity, better at capturing complex patterns*
Performance	*Struggles with complex patterns, lower accuracy*	*Better at capturing complex patterns, higher accuracy*

Results

The model training process showed significant improvement during the first 7 epochs, where training accuracy rose from 58.06% to 79.61%, and validation accuracy improved from 79.19% to 81.14%. After epoch 7, both training and validation accuracy plateaued around 82.28% and 82.56%, respectively, indicating that the model had largely converged. This suggests that while the model continued to improve, the gains were minimal after the 7th epoch, pointing to diminishing returns. The model avoided overfitting as validation accuracy remained stable, and training accuracy increased. Overall, training for 10 epochs is sufficient, with early stopping or learning rate adjustments potentially optimizing performance further without overfitting.

- **Accuracy:** The model achieved an accuracy of **83%** on the validation set, which is a good indication that it can generalize well to unseen data.
- **F1-Score:** The weighted F1-Score is **0.83**, indicating that the model has a balanced performance across the different classes. The F1-score considers both precision and recall, and this value suggests that the model performs consistently across both classes.
- **Precision:** The model achieved a precision of **0.83**, indicating that 83% of the positive predictions are correct. This is important for minimizing false positives.
- **Recall:** The model has a recall of **0.83**, meaning it correctly identifies 83% of the true positives. This is good as it shows that the model does not miss too many true instances.

Confusion Matrix



The confusion matrix reveals that the model performs well across all three classes, with the highest performance observed for **Class 2 (Mood Swing = 2)**, where both precision (0.83) and recall (0.84) are well-balanced. **Class 1 (Mood Swing = 1)** shows a strong recall of 0.85, capturing most true positives, but with slightly lower precision (0.81), indicating occasional misclassifications as class 1. **Class 0 (Mood Swing = 0)** has a precision of 0.84 and a recall of 0.77, suggesting that while the model is fairly accurate, it misses some instances of this class

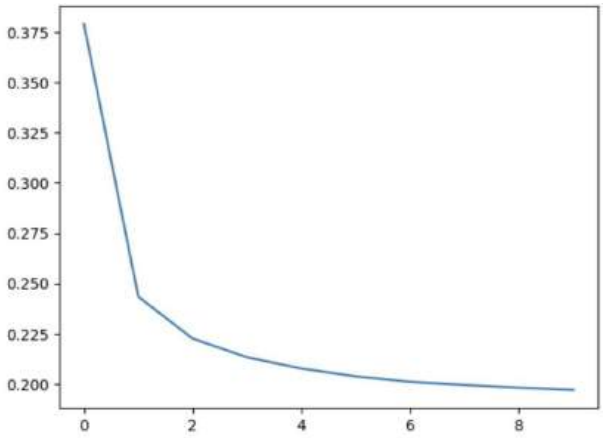
(higher false negatives). Overall, the model shows solid performance but could be fine-tuned to improve recall for class 0 without compromising performance for the other classes.

Classification Report

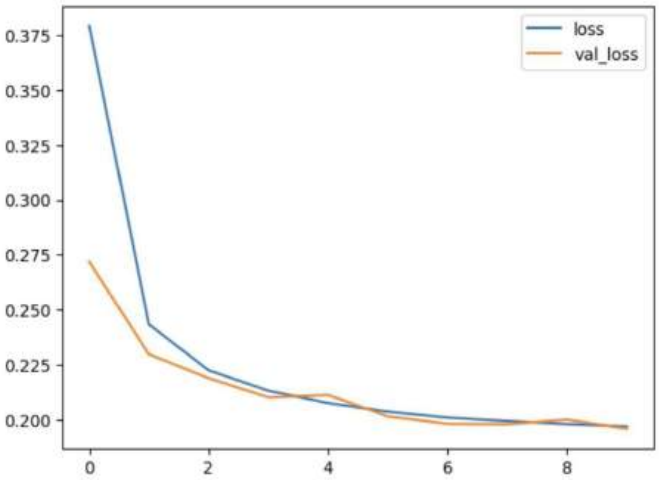
The classification report indicates that the model performs well across all three classes with a slight bias towards class 2.

- For class 0 (Mood Swing = 0), precision is **0.84** and recall is **0.77**.
- For class 1 (Mood Swing = 1), precision is **0.81** and recall is **0.85**.
- For class 2 (Mood Swing = 2), precision is **0.83** and recall is **0.84**. This shows the model is most accurate in predicting class 2.

The **training** and **validation loss curves** show a generally decreasing trend, indicating effective learning. Here are the key points:



Training Loss Curve: The training loss consistently decreases, indicating that the model is progressively learning and fitting the data.



Validation Loss Curve: The validation loss decreases and reaches a minimum of **0.2983** at epoch 9. After this, it fluctuates slightly, indicating that the model has reached near-optimal performance.

Sensitivity Analysis: The sensitivity analysis shows how the model's predicted probabilities for mood swing categories change as the input features are adjusted. As the 'mental health history' feature increases, the probability of mood swing category 0 (no mood swings) rises, indicating that a positive mental health history reduces the likelihood of mood swings, indicating a relationship between improved mental health history and the likelihood of mood swings. However, this relationship is more complex than a simple linear decrease and depends on how the other features influence the model.

Mental Health History	Mood Swing 0 Probability	Mood Swing 1 Probability	Mood Swing 2 Probability	Observation
0.0	0.7018	0.2434	0.0549	Lower probability for mood swing 0, higher for 1
0.1	0.8644	0.0869	0.0488	Increased probability for mood swing 0, decreased for 1
0.2	0.9464	0.0418	0.0118	Significant rise in mood swing 0 probability
0.3	0.9785	0.0204	0.0011	Further increase in mood swing 0 probability, almost no probability for 2
0.4	0.9811	0.0182	0.0008	Steady high probability for mood swing 0, low for 1 and 2

Discussion and Conclusion

In conclusion, this study highlights the potential of integrating AI-driven neural networks into mental health research to predict and address mood swings, a significant marker of psychiatric conditions. The neural network model demonstrates strong performance, with an accuracy of 83% and high scores for precision, recall, and F1-score. It shows balanced performance across all classes, though it slightly favors class 2. The model is successful at predicting mood swings based on the features provided, and the sensitivity analysis confirms that changes in inputs influence the model's predictions appropriately. Further improvements could focus on enhancing the model's ability to distinguish between classes 0 and 1, possibly by exploring advanced techniques such as hyperparameter tuning or adding more features.
These results underscore the promise of leveraging machine learning to advance mental health diagnostics and foster personalized care approaches to improve societal well-being.

Works Cited

Marwaha, Steven, et al. "Mood instability as a transdiagnostic marker of mental health problems." Journal of Clinical Psychiatry, vol. 75, no. 3, 2014, pp. 305–312.

Panchal, Nirmita, et al. "The Implications of COVID-19 for Mental Health and Substance Use." Kaiser Family Foundation, 21 Feb. 2021, www.kff.org/mental-health.

"COVID-19 Pandemic Triggers 25% Spike in Prevalence of Anxiety and Depression Worldwide." World Health Organization, 2 Mar. 2022, www.who.int/news-room.

"Adult Psychiatric Morbidity Survey." NHS Digital, 2014, digital.nhs.uk/data-and-information.

Kaggle. "Mental Health Dataset." Kaggle, www.kaggle.com/datasets.

Vahratian, Anjel, et al. "Symptoms of Anxiety or Depressive Disorder and Use of Mental Health Care Among Adults During the COVID-19 Pandemic — United States, August 2020–February 2021." Morbidity and Mortality Weekly Report (MMWR), vol. 70, no. 13, 2021, pp. 490–494. Centers for Disease Control and Prevention (CDC), www.cdc.gov/mmwr.

Patel, Vikram, et al. "Mental Health of Young People: A Global Public-Health Challenge." The Lancet, vol. 369, no. 9569, 2007, pp. 1302–1313.

Torjesen, Ingrid. "COVID-19: Mental Health Services Face ‘Tsunami’ of Cases as Lockdown Eases." BMJ, vol. 369, 2020, m1994, www.bmj.com/content/369/bmj.m1994.

American Psychological Association (APA). Stress in America: The State of Our Nation. American Psychological Association, 2017, www.apa.org/news/press/releases/stress/2017/state-nation.

World Health Organization (WHO). Depression and Other Common Mental Disorders: Global Health Estimates. World Health Organization, 2017, apps.who.int/iris/handle/10665/254610.

Mental Health Foundation. The COVID-19 Pandemic and Mental Health: Impact and Coping Strategies. Mental Health Foundation, 2021, www.mentalhealth.org.uk.

National Institute of Mental Health (NIMH). Mental Illness Statistics and Prevalence. National Institutes of Health, 2021, www.nimh.nih.gov.

Office for National Statistics (ONS). Coronavirus and Depression in Adults, Great Britain: June 2020. Office for National Statistics, 2020, www.ons.gov.uk.

Developmental Trajectories of Antisocial Personality Disorder During Childhood and Adolescence By Amina Gorman

Abstract

Antisocial Personality Disorder (ASPD) is a serious public health concern due to its increased risk for violent criminal behavior. It is crucial to properly understand the developmental trajectory of ASPD so that more preventative measures can be put in place to mitigate the effects the disorder has on quality of life for those afflicted. The purpose of this paper is to elucidate modifiable, socioenvironmental risk and protective factors of ASPD and to educate the public on its early intervention among youth with the disorder preceding ASPD, known as Conduct Disorder (CD). Findings from quantitative and qualitative studies from the United States were reviewed in this paper. Study findings reported parenting styles, social isolation, and negative peer affiliation as risk factors for CD. Parenting classes and classroom interventions were found to be protective factors for CD. Given the significance of adult intervention in child development, it is crucial to continue research in areas surrounding the betterment of the mental welfare of children, especially in view of the consequential impact adults have in shaping the personality development of the children they interact with.

Prevalence of ASPD

Antisocial Personality Disorder (ASPD) is a personality disorder characterized by behaviors such as a disregard for and violation of others' rights, failure to obey laws, deceitfulness, impulsiveness, physical aggression, and a lack of remorse. Between 3.9% and 5.8% of males and 0.5% to 1.9% of females in the US population meet the criteria for ASPD (Black et al., 2010, p. 2).

ASPD poses a significant public health concern because of its heightened risk for violent criminal behavior. In the past, surveys have shown up to 80% of male and 60% of female incarcerated offenders meeting the criteria for ASPD (Moran, 1999).More recently, a study indicates a reduced prevalence, with 37% of male and 26% of female incarcerated offenders meeting the criteria. However, prevalence estimates vary due to the varying populations in prisons and the methods in which the studies used to gather their data (Rotter et al., 2002). Studies have shown that adolescents may exhibit risk factors for the development of ASPD, such as impulsivity and juvenile delinquency, which may also predict premature mortality among incarcerated adolescents (Maurer et al., 2024). Given the significance of ASPD, it is crucial to understand potential risk and protective factors of ASPD during childhood and adolescence.

Developmental Course of ASPD

Candidates for an ASPD diagnosis must be 18 years of age or older with a diagnosis of conduct disorder (CD) in childhood or adolescence. The American Psychology Association defines CD as "a persistent pattern of behavior that involves violating the basic rights of others and ignoring age-appropriate social standards" (*APA Dictionary of Psychology*, 2023). Some of

the criteria that fall under this definition include aggressiveness towards people or animals, destruction of property, deceitfulness or theft, and serious violations of rules. A diagnosis of CD can be made when these behaviors significantly impair social, academic, or occupational functioning in children aged 15 or younger. CD has two subtypes: childhood onset, for children diagnosed before the age of 10, and adolescent onset, for children diagnosed between the ages of 10 and 15. Research indicates that individuals diagnosed with childhood-onset CD are more likely to display lifelong aggressive behaviors compared to those with adolescent-onset CD. Additionally, childhood onset CD increases one's risk of drug dependence and generalized anxiety disorder, as well as paranoid, schizoid, and avoidant personality disorders (Goldstein et al., 2006). Up to 50% of youth diagnosed with CD will meet the criteria for ASPD in adulthood (National Institute of Clinical Excellence, 2013).

The purpose of this paper is to elucidate modifiable, socioenvironmental risk and protective factors of ASPD to educate the public on its early intervention among youth with CD.

Methods

The criteria used for finding sources in this paper was as follows: the source must be a U.S sample written in English, published between 2000-2024, an empirical quantitative study, published in a peer reviewed journal, and must be related to one of the three major pathways being studied by this paper: the developmental course of ASPD, the risk factors, and the protective factors involved with ASPD.

Search terms used to find sources were: ASPD and/or mental health disorders, childhood CD, incarcerated offenders with ASPD, parenting styles, parenting, child abuse, CD to ASPD, teaching methods, cycle of abuse, socialization, risk factors, peer interactions.

Results

Risk Factors

The Effects of Parenting Styles

Parenting is an individualistic practice, with methods that vary across different cultural and socioeconomic backgrounds. Many of these more well-known methods have been studied for their effects on the development of children. Diana Baumrinds' four parenting styles: Permissive, Authoritarian, Authoritative, and Neglectful are some of the most recognized categorizations of parenting methods, but as more studies have been built around this topic, these definitions have altered with our general understanding. "Permissive" parenting is characterized as high levels of responsiveness coupled with low levels of control. There are very few boundaries set between the child and the parent, and the child is given most of the control in regard to their actions with little interference or punishment. Conversely, "authoritarian" parenting is characterized as high levels of control and low levels of responsiveness. Parents who follow authoritarian methods value discipline and have strict rules that limit the child's autonomy and are not known to be warm towards their children. "Authoritative" parenting combines

characteristics of the two previous styles, where parents have high levels of responsiveness and control. These parents are able to set limits and rules for their children, while still holding warmth and understanding towards them (Kawabata et al., 2011). A parenting style that was more recently added is "neglectful" parenting, where a parent has both low levels of control and responsiveness. These parents are unavailable in the lives of their children whether physically or emotionally (Howenstein et al., 2015).

ASPD has been known to have connections to "adverse childhood experiences" such as abuse and neglect (DeLisi et al., 2019). Physical child abuses' effect on the rates in which children develop ASPD has been supported by studies showing a significant increase. Families of children who develop ASPD often employ harsh and inconsistent disciplinary tactics, as well as generally being uninvolved in the child's life in a positive way, or in the supervision of the child (Patterson et al., 1993). These descriptions generally fall under a combination of authoritative and neglectful parenting styles according to Baumrinds' parenting styles.

A specific parenting practice that has been heavily researched in the past is the practice of coercion. Coercion occurs when a child behaves in a way that is undesirable to the parent, and the parent's response uses that instance of behavior to control the child (Smith et al., 2014). To elaborate, it develops into a positive feedback loop between parent and child aggression. Research demonstrates that children exhibiting symptoms of CD, such as callousness, combined with negative parenting practices, face an increased risk of becoming more antisocial in the future (Kochanska et al., 2013). As children become more callous and antisocial, parents will often increase their disciplinary measures, becoming more harsh. This cycle repeats itself until either party, though most commonly the parent, concedes "defeat" (Dishion & Snyder, 2016).

Additionally, parents who themselves have had difficulties regarding their conduct in their adolescence and childhood have been studied to be more likely to have children with similar traits. This is presumed to be due to the behaviors of the parent leading to impeded parenting practices that in turn exacerbate the potential of their children developing CD, or something similar (Raudino et al., 2013).

Social Isolation

There are many potential risk factors for ASPD and its severity. A lack of early socialization; that is, opportunities to have positive familial and peer relations, is a prominent risk factor. One example of this is ostracization from peers in childhood, which could potentially exacerbate the antisocial personality of a child (Holmes et al., 2001). The ability to self-regulate is an important developmental skill that affects a child's peer interactions (Calkins & Keane, 2009). Studies have shown that children with more aggressively prevalent CD are more likely to have difficulties with emotional regulation when compared to peers without or with less aggressive CD (Northover et al., 2015). Additionally, children with CD are known to be less likely to recognize the emotional state of their peers at a given moment (Martin-Key et al., 2018). If children are unable to gain experience in regulating their own emotions or in understanding and reacting accordingly to the emotions of their peers, their interactions moving

forward will likely be more negative, as the child could be seen as troublesome or "weird." This results in another feedback loop that exacerbates the child's isolation during adolescence.

Negative Peer Affiliations

The isolation of children from their peers in their developmental stages can lead to them either joining groups of similarly antisocial children or excluding themselves from socializing altogether (Holmes et al., 2001). The way a child is socialized can drastically affect the way in which they interact with the world. This is due to the fact that a large percentage of human behaviors are learned such as the fear of consequences (Waller et al., 2021). Socialization generally stems from the people a child spends their time with (Merriam-Webster, n.d.). This coupled with the idea that isolated children tend to bond with similarly isolated children can deepen how antisocial all children involved are. If these children are spending most of their school day with other children displaying CD behaviors, the likelihood of them developing more CD related tendencies from each other is high.

Protective Factors

A child's home environment influences many other aspects of their life, making stability at home essential. Parenting classes are one of the many ways parents can be involved in making a positive environment for their children. Studies show that often abuse is a cycle, going from one generation to the next, and that victims of child abuse are 40% more likely to become abusers themselves (Herzberger, 1990). One study discovered that a leading barrier for a parent with a history of child abuse preventing them from becoming a child abuser is dissociation (Narang & Contreras, 2005). Relying on the potential of dissociation is not a sustainable practice, nor is it a factor that can be controlled. Because of this, it is important for parents to find other preventative ways to end a cycle of abuse. Parenting classes are a great measure that can be taken to avoid continuing that cycle onto the next generation. Studies have shown that parenting classes such as Early Head Start can positively affect not only the language and methods parents use when communicating with their children, but they can also increase the cognitive stimulation in their children long term (Chang et al., 2009)

There are various approaches that potentially could lessen the possibility of CD developing into ASPD, many of which can happen within a child's immediate environment. One of these protective measures can occur in a child's school environment. School is one of the first places in which a child will interact with their peers, and teachers play a vital role in facilitating those early on interactions. As previously mentioned, these social interactions help to shape a child's understanding of the world around them, the expectations set for them by others, and even themselves (Grusec & Hastings, 2015). Schools are not only built with academics in mind as a function, but rather they help to socialize children to understand how society works. However, this process only truly works when teachers are all invested in the wellbeing of every one of their students.

Discussion

Goal of paper

This paper aims to enhance public awareness of the risk and protective factors associated with the development of ASPD during childhood and adolescence.

Results/Findings

This paper has uncovered the relevance of early childhood socialization in the development of CD, as well as ASPD. The people a child spends their developmental years watching are some of the most important people in that child's life, this includes parents, teachers, and peers. Social isolation severely diminishes the options the child has for socializing and, often because of this, isolated children tend to interact solely with other isolated children, causing the entire group to further isolate themselves from the rest of their peers. Without a variety of peers to interact with, these children will likely begin to exhibit the habits of the other children they interact with, habits that were likely the origins of the peer isolation, furthering the idea in their peers' minds that this child is "weird" or "mean" (Holmes et al., 2001).

Additionally, the methods parents use to interact with and discipline their children have a major effect on the development of that child. For children already experiencing CD symptoms, parenting methods that involve more aggression and control such as authoritarianism can lead to more push-back from the child and continue a cycle in which the parent has to compete with the child in some manner.

Trends

The studies discussed in this paper share many similarities regarding their findings. Something that is important to recognize from these findings is the significant role adults play in the developmental stages of a child's life, both in their home and school environments.

Interventions

Intervention methods for CD have been studied and should be recognized for their significance. One review paper has discussed in depth the method of therapy in a familial setting with the participation of the entire family (Woolfenden et al., 2001). This is a viable extension of what this paper discussed involving parenting classes, further connecting parents to the understanding that what they do impacts the livelihoods of their children significantly.

Strengths

A strength of this paper is its synthesis of the robust literature on parenting practices, family home environment, and developmental psychopathology. This paper additionally identified both the risk and protective factors of CD. While many studies have focused on risk factors for CD, fewer studies have identified protective factors due to the amalgamations of behaviors and personality traits that may contribute to CD. Thus, this review contributes to the

field by identifying protective factors of CD, which are more difficult to uncover in whole, leading to a reliance on a deeper knowledge of CD traits to make connections to protective factors involving specific CD behaviors (e.g., poor social skills, aggressive behavior). Another strength of this paper is its accessibility. This paper divulges background information to allow more public interaction with the material, thereby improving public awareness of the risk and protective factors of CD and ASPD.

Limitations

One limitation of this review paper was its inability to detect changes over time involving risk and protective factors. The importance of recognizing these changes may inform intervention development.

A limitation that was commonly seen in the review of papers revolving around CD was that many traits were not particular to only CD, but rather a larger group of callous-unemotional disorders in adolescents. An example of this was a study done on measuring impulsivity and its connection to CD, however, this paper discovered that there was not a noticeable difference in impulsivity levels between children with CD, and children who had other recognizable forms of psychopathy but did not present CD (Blair et al., 2020).

Future Directions of paper

In future review articles and potential future studies, the connection between the severity of CD and the environment the child grew up in should be further elucidated.

This information is important to properly divulge due to the long-term effects it can have on afflicted individuals, as well as those around them. Due to the connection to CD in childhood or adolescence, it is necessary for all adults who interact with or work with children to fully understand their role in the upbringing of children. It is equally, if not more so, important that these adults notice how their behaviors affect the children they are around.

Works Cited

APA Dictionary of Psychology. (2023, November 15). https://dictionary.apa.org/

Black, D. W., Gunter, T., Loveless, P., Allen, J., & Sieleni, B. (2010). Antisocial personality disorder in incarcerated offenders: Psychiatric comorbidity and quality of life. *Annals of Clinical Psychiatry: Official Journal of the American Academy of Clinical Psychiatrists*, *22*(2), 113–120.

Blair, R. J. R., Bashford-Largo, J., Zhang, R., Lukoff, J., Elowsky, J. S., Leibenluft, E., Hwang, S., Dobbertin, M., & Blair, K. S. (2020). Temporal Discounting Impulsivity and Its Association with Conduct Disorder and Irritability. *Journal of Child and Adolescent Psychopharmacology*, *30*(9), 542–548. https://doi.org/10.1089/cap.2020.0001

Calkins, S. D., & Keane, S. P. (2009). Developmental origins of early antisocial behavior. *Development and Psychopathology*, *21*(4), 1095. https://doi.org/10.1017/S095457940999006X

Chang, M., Park, B., & Kim, S. (2009). *Parenting Classes, Parenting Behavior, and Child Cognitive Development in Early Head Start: A Longitudinal Model*.

DeLisi, M., Drury, A. J., & Elbert, M. J. (2019). The etiology of antisocial personality disorder: The differential roles of adverse childhood experiences and childhood psychopathology. *Comprehensive Psychiatry*, *92*, 1–6. https://doi.org/10.1016/j.comppsych.2019.04.001

Dishion, T. J., & Snyder, J. J. (2016). *The Oxford Handbook of Coercive Relationship Dynamics*. Oxford University Press.

Goldstein, R. B., Grant, B. F., Ruan, W. J., Smith, S. M., & Saha, T. D. (2006). Antisocial Personality Disorder With Childhood- vs Adolescence-Onset Conduct Disorder: Results From the National Epidemiologic Survey on Alcohol and Related Conditions. *Journal of Nervous & Mental Disease*, *194*(9), 667–675. https://doi.org/10.1097/01.nmd.0000235762.82264.a1

Grusec, J. E., & Hastings, P. D. (2015). *Handbook of Socialization, Second Edition: Theory and Research—Google Books*. https://books.google.com/books?hl=en&lr=&id=P1-uCgAAQBAJ&oi=fnd&pg=PA251&dq=teachers+help+students+with+socialization&ots=NjENS_UpD9&sig=JPm-JkGtdIKqC1rANKfKiWjj6V0#v=onepage&q=teachers%20help%20students%20with%20socialization&f=false

Herzberger, S. D. (1990). The Cyclical Pattern of Child Abuse: A Study of Research Methodology. *American Behavioral Scientist*, *33*(5), 529–545. https://doi.org/10.1177/0002764290033005003

Holmes, S. E., Slaughter, J. R., & Kashani, J. (2001). *Risk Factors in Childhood That Lead to the Development of Conduct Disorder and Antisocial Personality Disorder*.

Howenstein, J., Kumar, A., Casamassimo, P. S., McTigue, D., Coury, D., & Yin, H. (2015). Correlating Parenting Styles with Child Behavior and Caries. *Pediatric Dentistry*, *37*(1), 59–64.

Kawabata, Y., Alink, L. R. A., Tseng, W.-L., van IJzendoorn, M. H., & Crick, N. R. (2011). Maternal and paternal parenting styles associated with relational aggression in children and adolescents: A conceptual analysis and meta-analytic review. *Developmental Review*, *31*(4), 240–278. https://doi.org/10.1016/j.dr.2011.08.001

Kochanska, G., Kim, S., Boldt, L. J., & Yoon, J. E. (2013). Children's callous‐unemotional traits

moderate links between their positive relationships with parents at preschool age and externalizing behavior problems at early school age. *Journal of Child Psychology and Psychiatry*, *54*(11), 1251–1260. https://doi.org/10.1111/jcpp.12084

Martin-Key, N. a., Graf, E. w., Adams, W. j., & Fairchild, G. (2018). Facial emotion recognition and eye movement behaviour in conduct disorder. *Journal of Child Psychology and Psychiatry*, *59*(3), 247–257. https://doi.org/10.1111/jcpp.12795

Maurer, J. M., Gullapalli, A. R., Milillo, M. M., Allen, C. H., Rodriguez, S. N., Edwards, B. G., Anderson, N. E., Harenski, C. L., & Kiehl, K. A. (2024). Adolescents with Elevated Psychopathic Traits are Associated with an Increased Risk for Premature Mortality. *Research on Child and Adolescent Psychopathology*. https://doi.org/10.1007/s10802-024-01233-6

Merriam-Webster. (n.d.). *Socialization Definition & Meaning—Merriam-Webster*. Retrieved November 25, 2024, from https://www.merriam-webster.com/dictionary/socialization

Moran, P. (1999). The epidemiology of antisocial personality disorder. *Social Psychiatry and Psychiatric Epidemiology*, *34*(5), 231–242. https://doi.org/10.1007/s001270050138

Narang, D. S., & Contreras, J. M. (2005). The relationships of dissociation and affective family environment with the intergenerational cycle of child abuse. *Child Abuse & Neglect*, *29*(6), 683–699. https://doi.org/10.1016/j.chiabu.2004.11.003

National Institute of Clinical Excellence. (2013). *Antisocial behaviour and conduct disorders in children and young people: Recognition and management*.

Northover, C., Thapar, A., Langley, K., & Van Goozen, S. (2015). Emotion Regulation in Adolescent Males with Attention-Deficit Hyperactivity Disorder: Testing the Effects of Comorbid Conduct Disorder. *Brain Sciences*, *5*(3), Article 3. https://doi.org/10.3390/brainsci5030369

Patterson, G. R., Debaryshe, B., & Ramsey, E. (1993). *A Developmental Perspective on Antisocial Behavior*.

Raudino, A., Fergusson, D. M., Woodward, L. J., & Horwood, L. J. (2013). The intergenerational transmission of conduct problems. *Social Psychiatry and Psychiatric Epidemiology*, *48*(3), 465–476. https://doi.org/10.1007/s00127-012-0547-0

Rotter, M., Steinbacher, M., & Smith, H. (2002). *Personality Disorders in Prison: Aren't They All Antisocial?*

Smith, J. D., Dishion, T. J., Shaw, D. S., Wilson, M. N., Winter, C. C., & Patterson, G. R. (2014). Coercive Family Process and Early-Onset Conduct Problems From Age 2 to School Entry. *Development and Psychopathology*, *26*(4 0 1), 917. https://doi.org/10.1017/S0954579414000169

Waller, R., Wagner, N. J., Flom, M., Ganiban, J., & Saudino, K. J. (2021). Fearlessness and low social affiliation as unique developmental precursors of callous-unemotional behaviors in preschoolers. *Psychological Medicine*, *51*(5), 777–785. https://doi.org/10.1017/S003329171900374X

Woolfenden, S., Williams, K. J., & Peat, J. (2001). Family and parenting interventions in children and adolescents with conduct disorder and delinquency aged 10-17. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.CD003015

The Impact of British Colonization on Indian Education By Alexa Bellavia

Research Abstract

This essay explores the impact of British colonization on Indian education and society, specifically from 1857 to 1914. The colonization of India by the British left significant and lasting implications on Indian society. By surveying recent Anglophone scholarly works on British colonial rule in India, this paper sheds light on the history of British educational reforms, the significance of policies such as changing the official language from to English, and erasing traditional education systems such as *Pathshalas, Tols, Madrasas,* and *Maktabs*. Furthermore, it explores how the British implemented a downward filtration method of education and the social impacts of such reforms. These social impacts include how the new education system privileged certain sections of society, worsened social divides, led to a loss of local culture, and catalyzed a crisis of identity amongst the locals.

British Colonization's Impact on Indian Education and Society

Education has impacted social dynamics, culture, and more throughout history in many different nations. In the case of India, under the reign of British colonization, education was taken over and used by the British as a tool for division and harm. The British needed to establish clear control of their Indian colony. As a white minority ruling over an Indigenous majority, they decided to use education as their tool for domination. Before British rule, the school systems in India, such as the *Pathshalas, Tols, Maktabs,* and *Madrasas* focused on developing their students' minds, bodies, and spirits.

I. Background

Along with teaching mathematics, language, and some science, the original Indian education systems included activities like yoga and focused on religion (Patel 73). These pieces of their education system helped keep students centered and led them to develop a moral code and skills based on conventional learning. When the British took over, they eliminated the focus on religion and spirituality, solely keeping the focus on subjects that would prove helpful regarding careers in government. This was because the British were using their education system to educate Natives in Western ways so they could act as intermediaries to the rest of the population, as the British could not reach them on their own. In doing this, they fostered extreme division among the Native population as they made their system of education available to only the wealthier upper classes. They used the resulting animosity to control the population since it prevented them from realizing the British were a common enemy.

One of the educational methods used in India before British colonization was known as the *Gurukul* system. This system was a residential school system that emphasized holistic learning and character development. All were considered equal at the *Gurukul*, and the *guru* (teacher) and the *shishya* (student) resided in the same house or lived near each other. The main focus was on imparting knowledge to students in an atmosphere of brotherhood, humanity, love,

and discipline. The curriculum focused on language, science, and mathematics through group discussion and self-learning. There was also a focus on arts, sports, crafts, and singing. These endeavors developed both intelligence and critical thinking. In addition, activities such as yoga and meditation generated peace of mind and physical fitness. It was believed that this curriculum helped personality development, self-confidence, self-discipline, intellect, and mindfulness necessary to life (Chandwani "The Importance of the Gurukul System"). The institution central to the elementary education system was the *Tol*, primarily to teach Hindu religious practices. These institutions were a community response to the needs of literacy and religious education of the children. Each institution was organized around a *guru*, and the students were supposed to spend considerable time learning Sanskrit (Patel 73).

Another institution that emerged was the *Pathshala* (school). This was a local elementary education center. The *Pathshala* curriculum was relatively secular: designed to teach language, basic mathematics, and skills related to agriculture and boat making. It was designed to meet the practical needs of the community. The *Pathshala* was established by a *guru* (teacher), who ran the center as a private enterprise. The *Pathshala* was open to students of all religions and castes. *Pathshalas* were predominantly attended by Hindu students, as the gurus were Hindu. A *Maktab* was an institution where Arabic or Persian was taught to children, mostly attended by Muslim students. *Madrasa* is an institution specially designed for Islamic education and culture. The *Madrasa* curriculum included grammar, mathematics, ethics, astronomy, literature, agriculture, medicine, logic, and government. This curriculum was taught in a way that encouraged students to think rather than learn by rote. During the medieval period, Muslim education was open to students of all faiths. Mughal sultans were interested in spreading knowledge and reforming the education system (Riaz 4; Patel 73). As a result, the number of *madrasas* increased significantly during the Mughal Empire (1556-1858).

The British take over of local education happened over a period of one-hundred years. The British East India Company was established in 1600 to control trade between the British and India. The battles of Plassey (1757) and Buxar (1764) laid the basis for the British conquest of India. Subsequently, the Court of Directors of the British East India Company articulated the Doctrine of Lapse as early as 1834. The Doctrine of Lapse was used to combat the issue of Hindu-state succession. According to this doctrine, if a monarch died without an heir and the monarch had allied or embraced Britain, the land reverted to the British. Also, if the British deemed the monarch incompetent, the land reverted to the British.

As a result of the Doctrine of Lapse, a large part of India came under the direct authority of the British East India Company; when British and other European traders arrived in India, they had to be in good standing with local rulers. This included the Mughal Empire. Although the East India Company was a "private venture", established by London merchants, its royal charter and militarily experienced employees gave the British a monopoly on trade in the region. The East India Company captured and built fortified trading outposts in port cities like Bombay, Madras, and Calcutta. A significant turning point in the East India Company's transformation from a profitable trading company to an empire came after the Battle of Plassey in 1757. The

Nawab, the leader of Bengal, one of India's wealthiest provinces, was upset with the East India Company for not paying taxes. The East India Company military leader in Bengal, Robert Clive, defeated the Nawab and assumed his wealth. This victory gave the East India Company broad taxation powers in Bengal (Gopal 110; Nogia, 1.2). As noted earlier, the Doctrine of Lapse allowed the East India Company to obtain control of a significant portion of India.

Once in control of these provinces, the East India Company could control the trade of goods produced there and maximize its profits by taxing Indigenous people. The Doctrine of Lapse also sparked the uprising of 1857, which united Indians across many classes, religious, and ethnic origins. This ended the East India Company's domain in India, transferring all the responsibility for administering India to the British Crown. The period of the British Raj lasted from 1857-1947. Over time, British rule emphasized control of private Indian enterprise in education rather than its development. The control took the form of a curriculum directed towards loyalty to the crown rather than a sense of patriotism (Ul Haq, et al. 423-24; Rahman, et al. 11).

Another example of British control over education was the narrow training for government employment rather than the spread of liberal education (i.e., more command over the English language rather than acquiring knowledge, skills, or values). As well as the dominance of the English language in administering trade, commerce, industry, and education (Motamedi 28-30).

As the British continued to take over the education system in India, people lost their culture, language, and even the possibility of a high-level education. British colonization and their commandeering of the Indian education system led to a loss of culture, increased division among social classes, and fostered animosity between different religious groups.

II. Educational Reforms Implemented by the British

The British introduced various educational reforms over the course of their reign in India, which greatly impacted culture, society, and other aspects of life for the Natives. One of the first educational reforms the British made was the Charter Act of 1813. The Charter Act of 1813 required the East India Company to develop a program of educators for the Indian people. The British formed a Committee of Public Instructors comprised of two opposing groups: the Anglicists, who believed in a filtration model of education with an English medium, and the Orientalists, who believed in introducing Western knowledge through the local language. (Nayak 178). The officials of the company considered the following three options:

(a) They could leave the indigenous education system as it was and provide state support for it.
(b) They could accept the indigenous system of education as the principal system but try to improve it by introducing Western knowledge through Sanskrit. This was the view of Orientalists.

(c) They could ignore the indigenous system and create a new system of education that would teach Western literature, philosophy, and science through the medium of English. This was the view of the Angliscists like Thomas Babington Macaulay (ibid).

Lord Thomas Babington Macaulay led the Anglicists. They believed it was the British government's duty to enlighten the people living outside of Europe. Lord Macaulay arrived in India in June 1834. He was a member of the Governor-General's Executive Council and was named as the President of the Committee of Public Instructors. In 1835, he was responsible for settling disputes between Orientalists and Anglicists (Nogia 1.4).

Lord Macaulay created his plan for education in India based on his Anglicist beliefs. This plan proved exceptionally influential as it was used as India's new British education model for some time. Macaulay's Minutes describes why the East India Company and the British government should invest in English language educators and promote European learning in India. While the Minutes noted the historical importance of Sanskrit and Arabic literature, it also claimed they had limitations. He further emphasized that:

a) The British government's primary goal should be to promote European literature and science among Indians and that "all funds appropriated for education would best be employed on English education alone."
b) The government funds should not be used to print oriental works.
c) All funds at the government's disposal would be spent on imparting to Indians a knowledge of English literature and science.

Thomas Babington Macaulay also stated his vision for an English education, "We must at present do our best to form a class who may be interpreters between us and the millions whom we govern, a class of persons, Indian in blood and color, but English in taste, opinions, morals, and intellect." (Gopal 112).

This shows that the British had one ultimate goal. It wasn't to improve Indian education or better the country but to train Natives to work for them and support their mercantilist agenda. As seen here, the British used these systems to create even greater division among the Indian people by capitalizing on this divide between the Western-educated and uneducated. They used education as a tool for division and control of the Indigenous masses.

Anglicists advocated for a filtration model of education. According to this model, English education was first imparted to the upper classes and would then filter down to the masses. As stated earlier, Macaulay's education aspired to "form a class who may be interpreters between us and the millions whom we govern, a class of persons, Indian in blood and color, but English in taste, opinions, morals, and intellect."(Motamedi 29; Ul Haq, et al.427; Gopal 112).

The main adversaries of the Anglicists were the British Orientalists. The Orientalists were East India Company officials, scholars, translators, and collectors who were in charge of learning and teaching Sanskrit, Arabic, and Persian in India. Orientalism was the Western

scholarly discipline of the 18th and 19th centuries that included the study of languages, literature, religions, philosophies, histories, art, and laws of Asian societies, especially ancient ones. Orientalism was also a school of thought among British colonial administrators and scholars who argued that India should be ruled according to its tradition and laws, opposing the "Anglicism" of those who argued that India should be ruled according to British traditions and laws. Despite this, Orientalism was still in favor of helping the British administration rather than acting in favor of the Indian people. Orientalists saw Indians as lower class and culturally backward compared to European societies. The British orientalists viewed the British as hardworking, honest, rational, and enlightened people. Conversely, Indians were portrayed as deceitful, irrational, and superstitious. Despite this Aryan view, the Orientalist believed that sufficient knowledge of Indian society was necessary to rule Indian people. They believed this information would be very useful in meeting the needs of the colonial state. It was believed by the Orientalists that this was the most efficient and profitable way to administer the Indian colony. The Orientalists were focused on primarily imperialist concerns despite their stance against Anglicism. (Borana 226).

The Anglicists eventually won the battle for education policies in 1835 when Lord William Bentnick made English the language of the courts and administrative offices (Gopal 112; Nogia 1.4-1.5). He also stated that the grand objective was to spread Western knowledge through English.

British policy towards India changed over time. Initially, the British showed no interest in India's education. The British believed any interference in India's educational matters might endanger its political and commercial enterprise. The British founded colleges to provide an oriental education to ensure its control of India and to keep both Hindus and Muslims from contesting British rule. Examples include the establishment of Alia Madrasa in 1780 by Warren Hastings (Governor-General from 1823-1828). In addition, Governor-General Lord Wellesley founded Lord Fort William College, which taught local languages to English officials (Rahman, et al. 7)..

The Governor-General of India, Lord Harding, issued a decree in October 1844 mandating English-speaking people should be given priority for all government jobs in India. This greatly increased English education in India, at the same time, hindering Indian culture.

The Filtration Model of education, utilized by the East India Company, was furthered by the model of education described by Sir Charles Wood in his Despatch of 1854. As the President of the Board of Control of the East India Company, Sir Charles Wood issued a Despatch to Lord Dalhousie, Governor-General of India, stated that a three-tier system of education be established. According to the Despatch, primary schools were to use the Indian language as their method of instruction, secondary schools were to use both the English and Indian language, and universities were to be established using English as their method of instruction (Gopal 113; Nayak 178).

The following occurred as a result of Wood's Despatch:

a) Bombay, Madras and Calcutta Universities were set up in 1857.

b) In all provinces, education departments were set up.
c) Bethune School was started for women's education.
d) British India saw rapid westernization of the education system with European headmasters and principals in schools and colleges.

Due to this Despatch, students at the secondary level were required to learn English before they could continue their education. The primary purpose of English education at the secondary level was to train the clerks and other lower-level officials required by the colonial administration. When English was made the official language, local people could only join the public service or hold office under the colonial government if they acquired some knowledge of English (ibid).

The educational reforms, instead of developing Indians, marginalized them. The major aim of the educational reforms were not to develop India but to have reformed subjects and intermediaries between the imperial power and the masses.

Overall, to modernize the educational system, the British opposed the Indigenous educational system and then introduced their system. Since this new system of education was entirely unrelated to Indigenous knowledge or education and foreign to Indian culture, it did not succeed long-term as it did in Western countries, nor did it help to develop a system of education in India.

III. The Impact of the British Model of Education on India

British colonial administrators implemented educational policies aimed at spreading Western knowledge and values among the Indian population. The introduction of English-medium education, modeled on British educational systems, created a new class of Western-educated elites who played pivotal roles in Indian society and politics. However, the emphasis on the English language and literature marginalized Indigenous languages and knowledge systems, perpetuating social inequalities and cultural dislocation.

Education in a country is closely related to its culture, as it provides transfers of intergenerational knowledge. Before the British arrival, India's education system was small in scale but well organized, with Muslim children being schooled in *Madrasas* and *Maktabs* and Hindu children being taught in *Pathshalas* and *Tols* (all referring to schools). These institutions taught children Arabic, Persian, Sanskrit, theology, grammar, logic, law, mathematics, metaphysics, medicine, and astrology (Ul Haq, et al. 424; Motamedi 28-32). The British government, however, ignored this faith-based education system and replaced it with a British system (Motamedi 25-32; Patel 3; Ul Haq, et al. 424; Gopal 112). This action affirms a colonial motive the British government intended to fulfill by introducing English education into India.

English education was not universal; it was designed only for the British and Indian elites. Individuals from poorer backgrounds were not privileged to participate in this education project. People who received English education viewed themselves as superior to those not educated under this system. As such, the education system divided people into two distinct

classes: a class that received a British education and a class that was deprived of that education. Consequently, this education system encouraged class distinction and engendered antagonism among India's natives, thus weakening cohesion and contributing toward fulfilling the British vision of domination and exploitation.

In their attempts to develop their colonies, the English employed a system of education with two primary goals:

a) To westernize the urban areas
b) To urbanize the rural areas (Motamedi 28-30)

As part of their modernization and educational plan, the British spent money exclusively in cities and/or colleges where loyal servants and clerks were trained to administer the colonial government. They established universities in Bombay, Calcutta, and Madras (Motamedi 30).

The English language versus vernacular education created a divide between the elite and the masses, as English was a means of entry to the elite class. As such, English education attracted intellectually superior individuals from rural areas to metropolitan centers. The schools that provided English education were detached from Indigenous cultures in both languages and social values. These schools did not hold out the prospect of reintegration into Indigenous cultures to its students. This method of education allowed for the development of an elitist philosophy- all that is rural is bad, all that is urban is better, and all that is foreign is best (Rahman, et al. 8; Motamedi 30).

The British imposed the English language and a dependency on the English education model in India. Currently, the English language is still predominantly used by the elite and the educated in India.

The English education in India dominates Indian culture and values. The educational policy that the British imposed on India was partly responsible for creating rural neglect and advancing urban areas. Under British rule, urban areas were developed for the benefit of colonial power at the expense of underdeveloped rural areas. The underdeveloped rural areas essentially were colonies of the developing urban centers and productive agricultural areas (Motamedi 31). The exploitation of raw materials and labor in underdeveloped regions allowed the growth of the modern sectors. Thus, developing urban centers and productive agricultural areas is at the expense of under-development of rural areas.

The British's expansion of education and educational facilities in India did not produce a literate population. Rather, the British eradicated the Indigenous system of education, which emphasized the English language as its focus in the curriculum. The British emphasized the growth and development of an elite class in urban areas and neglected rural areas.

In the centuries of colonial domination. English-educated elites emerged as leaders in the overall social structure of India. During British rule, a minority of the population of India received a significant amount of support in the sphere of education, which enabled them to

secure important positions in both administrative and commercial sectors (Di Bona 618; Rahman, et al. 8; Motamedi 101).

These English-educated elite groups led the campaign for self-government and independence later on. However, these leaders needed to restructure the education system and economy to provide educational opportunities and jobs to people living in underdeveloped rural areas. Rather, emphasis was placed on the growth and development of that elite class at the expense of the rural masses. (ibid).

It is important that one acknowledges the cultural values of the Indigenous population and develops a system that addresses the specific issues of cultural resistance and acceptance rather than resorting to the wholesale importation of a Western model of education. The British failed to accomplish this in India.

Colonial educational institutions sought to convince Indian children of the superiority of the British. The colonists wanted to educate the Indians just enough to work for their requirements. They did not encourage an analytical thought process. Indians were taught to believe that they were both primitive and ignorant. Despite being a seat of glorious achievements in knowledge, architecture, and arts, India was reverted by the British into a poor importer due to ruthless economic exploitation. English education cultivated the idea that the British were superior to Indians. English education in India was focused on "general subjects" instead of scientific and technological education. The neglect of these areas of study created a gap between the ruling power and its colonies. This contributed to the intellectual reliance on the ruling power of the colony.

As English became the language of all formal sectors of the state, including education, trade, commerce, and government, native Indians started to neglect their languages. In addition, traditional education models were abolished, and the major emphasis was placed on acquiring a modern Western education. This meant traditional subjects like Sanskrit philosophy and classical arts, which were integral to the indigenous system of education, were neglected. The Indigenous system of education not only imparted academic knowledge but also taught moral values, cultural traditions, and spiritual practices. Western education undermined these aspects of Indigenous education, leading to a loss of cultural identity. (Patel 3).

Successive generations faced an identity crisis, torn between Western influences and their traditional roots. This has often led to a lack of pride in one's heritage and a preference for Western lifestyles and values. (Motamedi 101-102).

The British heightened tensions between Hindus and Muslims when it abolished Persian and adopted English as the official language in India in 1835. The measure was also followed by the introduction of English in schools supported by the East India Company, replacing Persian and Sanskrit. Both these steps benefited Hindus and disadvantaged Muslims primarily because of two reasons:

a) Hindus had already been learning English and there was already a significant section of the Hindu elite who were well versed in English.
b) Muslims thought it was against their religion to learn English.

The replacement of Persian with English as the official language resulted in a huge loss for the Muslims. It resulted in a significant loss of employment for Muslims in government service, also diminishing their chances of finding government employment in the future. (Ul Haq, et al. 427).

Muslims suffered economically, socially, and politically as they resisted the new system of education. An annual report of 1852 regarding the progress of education in Bengal observed that the Muslim community was way behind the Hindu community. In 1871, W.W. Hunter published "The Indian Musalmans," according to which the number of Muslim government officials had declined for the reason that they lacked modern education. Thus, British rule impacted Hindu and Muslim communities differently. For Hindus, British rule was seen as a change of rulers (from Muslims to British). However, for Muslims, it was also a loss of their power.

Consequently, a majority of Muslims resisted the British colonial empire, and when the modern reforms were introduced, many of them mistrusted the reforms. So, under British colonialism, the economic, educational, and political situation of the Indian Muslims was adversely influenced. As a result, Muslims were politically and economically alienated from mainstream regional affairs (Ul Haq, et al. 427).

Despite being alienated from mainstream regional affairs, this did not mean all Muslims were completely complicit. Sir Sayyid Ahmad Khan sought to change the view of Muslims toward Western education to improve prospects for Muslims in British colonial India. He was convinced that the best of Western Civilization could and should be assimilated by Muslims because the pure Islam taught by the *Qur'an* was simply not opposed to Western civilization but was, in fact, its ultimate source (Parray 163). His interpretation of Islam was guided by his belief that Islam was compatible with reason and the laws of nature and, therefore, in perfect harmony with modern scientific thought. He argued that Islam's teachings concerning God, the Prophet, and the *Qur'an* are compatible with modern science, which involves the discovery of the work of God through natural laws. In other words, Sir Sayyid argued that Islam is "in full correspondence with reason." (Parray 164). Further, he equated reason with understanding and considered it an acquired quality that enables humans to distinguish between good and evil, right and wrong, and proper and improper. According to Sir Sayyid, who used terms like understanding, reason, and intellect interchangeably, the only criterion for a person having reason, intellect, or understanding is behavioral rather than substantive.

In his reaction to British Imperialism and its impact on Indian society, Sir Sayyid was the first Indian Muslim to feel the need for a fresh orientation of Islam and worked for it.

A contrasting point of view was the Deoband Movement was established by Mohammed Qasim Nanautavi and Rashid Gangohi as a revivalist movement. Instead of essentially working in tandem with the British, they attempted to work against them. This movement had two objectives:

a) Propagating pure teachings of the *Qur'an* and *Hadith* (teachings of Muhammed) among Muslims

b) Keeping alive the spirit of *jihad* (holy war) against foreign rule. (Parray 163; Kumar 43)

Despite all of this, Muslims and Hindus were still completely ostracized from one another due to the education policies and their impact on the other. This was all part of Britain's divide-and-rule plan so that the Muslims and Hindus would not unite against the British throne.

IV. Conclusion

Britain's education policy in India was not only an instrument of domination but also a weapon of oppression. An effective education system must obtain the consent and participation of learners, teachers, and administrators. The British ignored this concept and made no effort to ascertain what Indian students needed from their education. The British, as justification for stripping India of its wealth, viewed their system of education as superior, and to develop India, they introduced the Western education system to them. Rather than develop India, in reality, their education system was designed to manufacture skilled clerks to help them administer the colony and maximize revenue for the British. While doing this, the Western education system marginalized indigenous knowledge and undermined the cultural identity of Indians.

The *Gurukul* system, the foundation of Indian education for centuries, was systematically marginalized. Traditional subjects such as Sanskrit, philosophy, and classical arts, integral to the *Gurukul* curriculum, were neglected. The implementation of Macaulay's education policy led to an erosion of India's rich cultural heritage. The *Gurukul* system provided more than academic knowledge. It also taught moral values, cultural traditions, and spiritual practices. The British policy on education, with its focus on Western ideals, undermined these teachings, leading to a loss of cultural identity among the educated classes.

The essential teachings of the *Gurukul* system were taught through group discussions and self-learning. There was also a focus on arts, sports, crafts, and singing that developed intelligence and critical thinking. Activities such as yoga and meditation generated positivity and peace of mind. Each student was required to perform daily chores on their own to learn practical skills. This education led to personality development, self-confidence, self-discipline, intellect, and mindfulness which is necessary to navigate one's life.

The focus under British rule shifted from holistic education to note-learning and examination-oriented studies. The comprehensive development of students, which was the hallmark of the *Gurukul* system, gave way to a narrow, utilitarian approach to education. This narrow approach to education was criticized by the Indian National Congress in its 1806 Resolution on National Education.

A significant section of Indian society, particularly the elite, became heavily Westernized. This created a socio-cultural divide between those educated in English and those who continued to follow traditional ways. This divide persists to this day.

In the centuries of colonial domination, English-educated elites emerged pre-eminent in the overall social structure of India. During British rule, a minority of the population of India

received a large measure of support in the realm of education, which allowed them to secure important administrative and commercial positions. It was the educated elites who were at the forefront of the campaigns for self-government and independence. However, when the elite class came to power, they failed to reform the education system. Rather, emphasis was placed on the domination of one sector of the population over another and the growth and development of that elite class at the expense of the rural masses.

The English-educated elite had better access to economic opportunities, exacerbating social and economic disparities. Those who were not proficient in English found themselves at a disadvantage, perpetuating cycles of poverty and exclusion. This perpetuated the education system's alignment with Western standards and the dependency on Western knowledge. This has resulted in successive generations facing an identity crisis. South Asians are torn between Western influences and traditional roots. This has often led to a lack of pride in one's heritage and a preference for Western lifestyles and values.

To conclude, British educational policies implemented in colonial India led to social division, animosity between different religious groups, and an overall loss of culture. The British were able to turn a powerful tool like education into a deadly weapon that separated people and economically hurt many as well. Colonial history is a model that shows how people have been treated poorly and accurately depicts how they should not be treated in the future. Education is powerful and can change lives for the better or worse.

Works Cited

Borana, Mihir. “Orientalism: Nature and its impact on Indian history”. *academia.edu.*

Chandwani, Nikhil. “The Importance of the Gurukul System and why Indian Education needs it”. *The Times of India.* 2019.

Di Bona, et al. “The Development of Educational Underdevelopment in India”. *In Asia Profile.* 1977.

Ganapati Nayak, Jyoti. “Historical Review of Educational Reform in India for the 21st Century”. *JETIR.* 2019.

Gopal, N.R. “The Implications of British Colonial Domination on the Indian Cultural Ethos”. *UAEM.* 2021.

Kumar, Sanjeev. “Socio-religious reform movements in British colonial India”. *International Journal of History*. 2020.

Motamedi, Vahid. “Consequences of the British Model of Education in Colonized Third World Nations with Special Reference to India”. 1994.

Nogia, Harshita. “British Colonial Domination’s Effects on Indian Cultural Values”. *IARI.* 2015.

Parray, Tauseef Ahmad. “Muslim Responses to Imperialism in India: A study of the Educational Reforms of Sir Sayyid Ahmad Khan”. *International Journal of History.* 2012.

Patel, Lalitbhai. “Gurukul education system of ancient India and Indian Education Policy Historical practice of 1947-2019 A.D. *International Journal of History.* 2021.

Riaz, Ali. “Madrasa Education in Pre-colonial and Colonial South Asia”. 2011.

Rahman, Aziz, et al. “The British Art of Colonialism in India: Subjugation and Division”. *Nova Southeastern University Libraries.* 2018.

Ul Haq, et al. “An Analysis of the 19th Century Educational Reforms Of the Subcontinent through the Postcolonial Lens”. *Pakistan Journal of Social Research.* 2022.

Utilization of Whole-Genome Sequencing as an Advanced Detection Tool for Cutaneous T-Cell Lymphoma By Kathya Sareddy

Abstract

Cutaneous T-cell lymphoma (CTCL) is a slow-developing cancer that starts in T-cells and invades the skin. It is an extremely challenging cancer to detect or diagnose, as cancer cells do not appear in lymph nodes until Stage 4, and lymph nodes do not enlarge until Stage 3 or 4, allowing the disease to potentially progress beyond the point of effective treatment. Current diagnostic techniques have traditionally relied on biopsies and protein tests. Advances in genetic testing technology have the potential to precisely diagnose CTCL. Previous research on CTCL has linked mutations in many genes, particularly those associated with the immune system pathways, JAK-STAT and NF-kB. I hypothesize that those pathways get affected, leading to a dysregulated cell cycle, which in turn causes cancer. We will explore how whole genome and single locus sequencing can serve as a diagnostic tool for CTCL. Applying genetic testing to diagnose CTCL could lead to earlier detection, increased treatment success rates, and serve to identify targets for precision medicine.

Summary

We explore how whole genome and single locus sequencing can serve as a diagnostic tool for CTCL, a cancer that is hard to detect until later stages.

Introduction

Cutaneous T-cell lymphoma (CTCL) is a type of non-Hodgkin's lymphoma that affects T-cells, which subsequently infiltrate the skin (Bagherani et. al, 2016). This is important, as this is a different mechanism than other skin cancers like melanoma. There are two main forms of CTCL, mycosis fungoides, which is the most common at 60% of cases, and Sézary syndrome, which comprises <5% of cases (Hague et. al, 2022). The main symptoms of mycosis fungoides are dry, red, scaly patches of skin. In later stages, tumors develop, metastasize, and cancer cells are found in the skin. In Sézary syndrome, a large rash covers most of the body and a low amount of cancerous Sézary cells develop in the blood. In the last stages, the Sézary cell count is high and the lymph nodes become enlarged. The cancer is also more likely to spread to organs like the liver or spleen (Markman, 2022; Healy, 2024).

In both cases, the cancer originally presents as common dermatological conditions such as eczema or psoriasis. Even in later stages, the cells themselves do not look too abnormal (Healy, 2024). Because of this, early testing for CTCL is difficult and inaccurate. Our current diagnostic is a skin biopsy, which can delay the discovery of the cancer until later stages (Markman, 2022). Due to the relative rarity of CTCL (accounting for only 4% of non-Hodgkin's lymphoma cases), there has been limited development in treatment options for this cancer (Wiese et al., 2023). However, the survival rates and median overall survival from earlier stages (70-88% and 21.5-35.5 years, respectively) are much higher compared to later stages (15-18%

and 1.4-3.8 years) (Hristov et al., 2023). Diagnosing this cancer earlier would lead to better treatment success rates, a crucial need for the 12.4 million patients diagnosed in the United States annually (Wiese et al., 2023).

Some risk factors of CTCL are immune system weakening through AIDS/HIV or drug suppressants, exposure to other chemicals, and bacterial or viral infections (Markman, 2022). In a study published on multiple pairs of twins to explore the heredity of CTCL, it was found that the female to male ratio is equal to 1:1.8 and CTCL commonly developed around the average age of 53 (Odum et. al, 2017). African Americans are also twice more at risk than other racial groups (Histrov et al., 2023). Treatment for CTCL can include chemotherapy, radiation therapy, photodynamic therapy, immune therapy, targeted therapy, and other medicines. Current diagnostic and staging procedures are a complete physical exam, multiple biopsies, blood tests, and looking through medical history (Markman, 2022). Most commonly, a complete blood count is performed. CTCL patients are more likely to have elevated white blood counts (Brown et al., 2024). Sézary blood cells can be specifically tested for with flow cytometry, as their shape is abnormal due to the cell's nucleus and they are larger than a regular blood cell (**Fig. 1**) (Naiem et al., 2013).

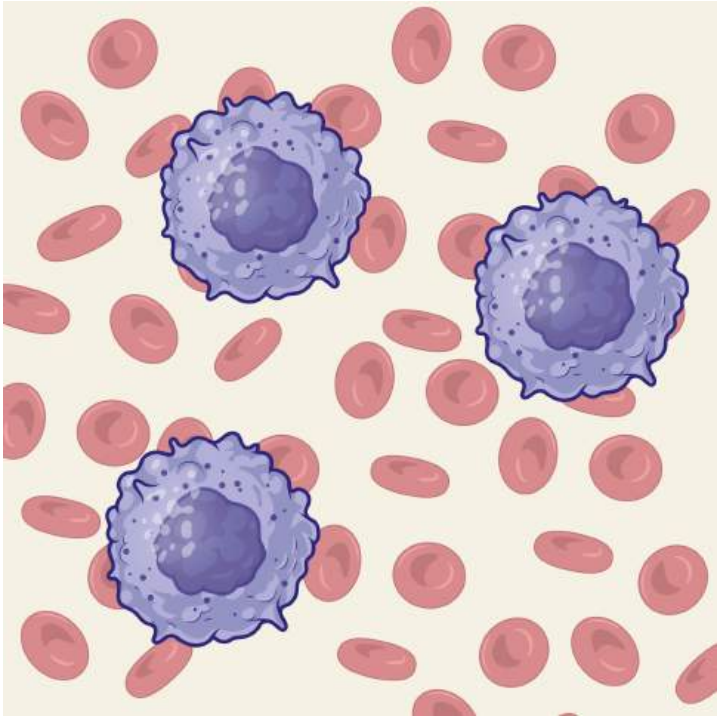


Figure 1. Sézary cells in the bloodstream
Created in BioRender. Ho, L. (2024) https://BioRender.com/m58o411

Serum protein electrophoresis is performed to test for abnormal protein levels, which are common in lymphomas. If protein levels are abnormal, a bone marrow biopsy is recommended (Shinohara, 2024). Skin biopsies are done to test for cancerous T-lymphocyte cells. Lymph node biopsies are done to confirm the diagnosis and spread of the cancer (Horwitz et al., 2019). These

are not as effective, because these biopsies usually don't show cancer cells (Ludmann et al., 2023). To find the cancer in the early stages, we need better diagnostic tests.

Genetic testing is a great diagnostic tool that is being used in the detection of many diseases and conditions such as Down syndrome, trisomy 18, trisomy 13, neural tube defects, cystic fibrosis, sickle cell anemia, Fragile X syndrome, and Tay-Sachs (Griffin et al., 2023). New research has shown that genetic testing can help detect certain cancers, such as breast and colon cancers (Winstead, 2023). The development of new genetic testing technologies and recent discoveries in the nature of the disease could make CTCL a candidate for genetic diagnostics.

Genetic Testing Methodologies

A classic method of genetic testing that has been used since the 1840s is cytogenetics. Cytogenetics is the study of chromosomes, their behavior during cell reproduction (mitosis and meiosis), their influence on phenotypes, and their structural changes. Some early discoveries in cytogenetics are the chromosome theory of inheritance, the identification of 46 chromosomes in humans, the discovery of the X and Y chromosomes, and karyotypes. The idea of clinical cytogenetics was made in the 20th century when Jerome Lejeune proved that chromosome abnormalities affect the phenotype (Mathew et al., 2024). Cytogenetic testing originally started with banding techniques, such as quinacrine banding (Q-banding), constitutive heterochromatin banding (C-banding), reverse banding (R-banding), nucleolar organizing regions banding (NOR-banding), and Giemsa-banding (G-banding). The Q-banding technique treats the chromosome with quinacrine to create a banding pattern of alternating dull and bright regions. The highly fluorescent areas are filled with adenine and thymine (Burnett et al., 2006). A big milestone for cytogenetics was fluorescent in-situ hybridization (FISH), which is a technique of mixing fluorescent probes to detect certain nucleic acid sequences, and it is still commonly used today. Another common technique we progressed to is molecular array comparative genomic hybridization or CGH, which compares DNA strands and detects chromosomal changes like deletions, duplications, and amplifications (Kannan et al., 2009; Moore et al., 2021). Some diseases discovered through cytogenetics are Down syndrome (trisomy 21) and Edwards syndrome (trisomy 18) (Ataman et al., 2012; Balasundaram et. al, 2023). Due to the progression of genetics in the past decades, there are currently many different methods for the diagnosis of genetic diseases, but the main categories are biochemical and genetic testing.

Rather than identifying DNA mutations, biochemical testing is used to identify defects in proteins and other molecules. Some examples of biochemical tests are antibody-antigen testing, total protein level tests, high-performance liquid chromatography (HPLC), and mass spectrometry. Antigens are anything that triggers an immune response, and looking for those through tests can help see if the immune system is stimulated and if an antigen is active. Antibody tests look for past times the immune system has been stimulated, and those tests are most commonly used for detecting past infections or diseases. These types of tests are used to diagnose respiratory infections such as influenza, respiratory syncytial virus (RSV), and coronavirus disease (COVID-19), among many other diseases. Fluid samples are typically used

and mixed with a reagent to get the results in 15-30 minutes (Zimlich et al., 2024). In total protein tests, blood samples are collected and the serum is isolated from the red and white blood cells. Then, protein electrophoresis is performed on the serum to separate the proteins by size. Total protein tests can detect many things, such as kidney damage, liver disease, malnutrition, edema, autoimmune disease, cirrhosis, and some cancers like multiple myeloma (Martel et al., 2018). High-performance liquid chromatography (HPLC) is a way of separating liquid samples. The sample is pushed by a high-pressure pump through a column. The different solubilities of each component of the sample lead to different velocities at which they move through the column, leading to separation. The results are measured by a detector connected to the column and the data is input to make a chromatogram (Bower, 2024). HPLC is primarily used for testing concentrations in medicine and can detect abnormal hemoglobin variants like sickle cell or thalassemia (George et al., 2001). Mass spectrometry converts molecules into ions and those ions are then sorted by their mass-to-charge ratio by magnetic and electric field manipulation. Those ratios are measured with a detector and the results are on a mass spectrum chart. Mass spectrometry is commonly used to diagnose diseases through biomarkers, which are molecules found to be associated with certain pathways, cellular processes, etc. (Garg et al., 2023). Overall, these and other biochemical tests are very fundamental for diagnosis and are used for clinical applications. The higher level ones such as HPLC and mass spectrometry are less commonly used and more expensive but are still essential in pharmaceutical uses and drug testing.

The main kinds of genetic testing are Sanger sequencing and next-generation sequencing (NGS). These types of tests are based on polymerase-chain reactions (PCR) and how it can target sections of DNA, even though the whole genome is large. PCR has three main steps: denaturing, annealing, and elongation. First, the DNA's double helix structure denatures due to heat. Then, designed DNA primers attach to a complementary section of the parent strand DNA through hybridization (annealing). Lastly, the targeted DNA sequence is duplicated repeatedly with the help of the primers, nucleotides, and DNA polymerase (which works from the 5' end to the 3' end), to create an elongated version of the targeted DNA (Smith, 2024).

The idea of targeting certain genes, which is evident in PCR, is used in single-gene sequencing, specifically Sanger sequencing. To determine the position of specific nucleotides in a sequence, Sanger sequencing employs dideoxynucleotide triphosphates (ddNTPs).The difference between ddNTPs and regular DNA nucleotides is the missing oxygen on the 3' end, preventing the phosphodiester bond from forming. This stops the strand of DNA from building any further. There are four separate ddNTPs, one for each of the four different nitrogenous bases (nucleotides), and they are marked with fluorescent proteins. The first part of the process of Sanger sequencing is a PCR, denaturing the double helix, and then cooling it back down, and adding nucleotides with primers and DNA polymerase. The fluorescent ddNTPs are also incorporated with the primers and added by the DNA polymerase with their complementary bases, which then prevents elongation by DNA polymerase. This creates DNA fragments of different lengths, which all have a fluorescent ddNTP at the end. Then, the fragments can be separated by gel electrophoresis, and a laser excites the fluorophores in the ddNTPs. The

smallest fragment is simply the 5’ end of the original DNA strand. The DNA sequence is determined by organizing the fragments from the smallest fragment (5’ end) to the longest fragment (3’ end) (**Figure 2**) (Heather et al., 2016). Sanger sequencing is considered the most accurate type of DNA sequencing to date, with a high success rate of 99.99% (Shendure et al., 2008).

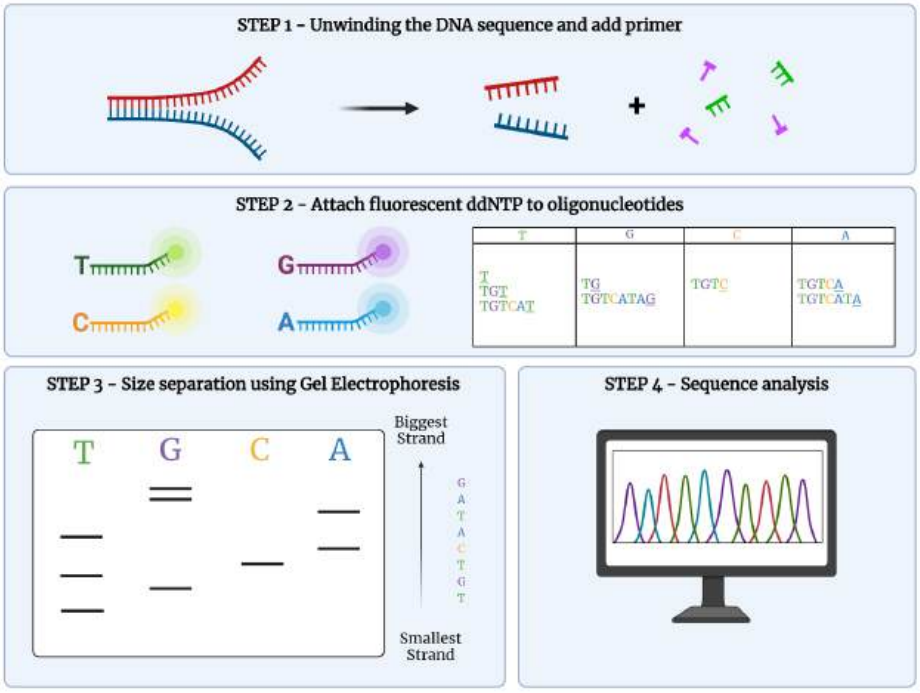


Figure 2. Sanger sequencing process
Created in BioRender. Ho, L. (2024) https://BioRender.com/m58o411

Sanger sequencing is mainly used for figuring out a single gene, or single locus sequencing, but is also used for diagnosing diseases. It can test for hereditary conditions, specifically autosomal recessive conditions (i.e. cystic fibrosis). It can also act as a foresight for testing for specific genes, for example, breast cancer and the *BRCA1* gene variant. Currently, some machines conduct Sanger sequencing and read the bands from the gel electrophoresis, making the process faster. The main limitation of Sanger sequencing is sequencing multiple genes or many samples. It is not cost-efficient and it has a limited input of DNA. To solve these limitations, researchers developed new technologies collectively labeled next-generation sequencing (Frost et al., 2022).

Next-generation sequencing (NGS) is a technique that enables the simultaneous analysis of millions of base pairs and DNA strands. NGS has many variations that use different methods for library prep, PCR amplification, and base reading. However, the main method for NGS is fragmenting the DNA, then marking those fragments with some sort of indicator (library prep), putting those fragments through an NGS sequencer (sequencing), and then analyzing and comparing the given data to a reference genome (Qin, 2019). There are two main kinds of NGS that are typically used in clinical testing: whole genome sequencing (WGS) and exon

sequencing. WGS analyzes all of the patient's DNA to determine the entire genome sequence. This means that WGS can detect any mutations in the DNA, which is why it is currently used to diagnose genetic disorders and developmental delays. It's also used for newborn screenings and cancer detections and is a breakthrough for personalized medicine (Brlek et al., 2024). Exon sequencing analyzes just the coding regions (exons), specific sections of DNA that are transcribed to RNA, and then translated to proteins. Only 1% of the human genome is exons, and exon sequencing is a much more efficient way to identify disease-causing mutations, as exons are the only pieces of DNA that produce primary proteins (Brlek et al., 2024). Exon sequencing has been used in the detection of diseases such as autism, epilepsy, brain malformations, congenital heart disease, and neurodevelopmental disabilities, and has identified new disease pathways (Retterer et al., 2016). Overall, NGS can pinpoint specific mutations that can lead to the discovery of a disease-causing mutation.

Analogously, NGS could be used for CTCL, as much research points towards multiple gene mutations being linked to causing CTCL. Numerous studies show that CTCL could be a cancer that is caused by genetic mutations, which can be identified easily with NGS. Biochemical tests can also be applied to test for certain proteins in patients and see if it is linked to the genes mutated in NGS. NGS is a good alternative for current diagnostics and could solve the problem we face when testing for CTCL.

Current Understanding of the Genetic Causes of CTCL (Results)

Genetic diseases are currently identified through sequencing because there is typically one gene that leads to a direct association with the disease.An example of this is cystic fibrosis, where a mutation in the CFTR gene leads to the dysfunctional transport of salt and water across epithelial cells. This creates the mucus around the lungs, which then leads to cystic fibrosis (Csanády et. al, 2018). There are direct-to-consumer (DTC) tests for diseases that have clear markers. However, using genetic testing or DTC tests for cancers to pinpoint the specific gene(s) with cancerous mutations is a challenge. For CTCL, this remains an issue as studies have found a varied number of mutations linked with this cancer. Using NGS to identify a panel of genetic mutations is the best way to figure out each of these mutations and compare them to multiple patients, leading to an effective diagnosis of CTCL.

The majority of these genes are associated with T-cell activation and apoptosis, chromatin remodeling, DNA damage response, and NFκB and JAK/STAT signaling pathways (Choi et al., 2015; Litvinov et al., 2017; Argyropoulos et al., 2020; Lai et al., 2021; Zhang et al., 2021; Park et al., 2021; Motamedi et al., 2021; Song et al., 2022; Tensen et al., 2022; Ren et al., 2023). One study found seventeen genes linked with the cancer through whole exon sequencing (WES), whole genome sequencing (WGS), and RNA sequencing on forty CTCL patients, all of them in late stages (IVA1 - IVB). Scientists have hypothesized that cancer targets the immune system pathways NF-κB and JAK-STAT along with multiple tumor-suppressing mechanisms (Choi et al., 2015). A different study found eighty-six putative genes by incorporating SSNVs (somatic single nucleotide variations), SCNVs (somatic copy number variation), and structural variant

data (Park et al., 2021). Another study found a hundred and fifty recurrent genes and identified fifteen genes associated with the progression of mycosis fungoides by using data from other publications. Those genes are correlated with cell proliferation, immune checkpoints, apoptosis, and immune response (Motamedi et al., 2021). In Sézary syndrome, it is hypothesized that five major processes are affected: cell cycle regulation, MAP-kinase signaling, T-cell receptor/NF-κB signaling, JAK-STAT signaling, and chromatin modification. The data is from multiple studies and research that took pools of patients–one study had as many as 220 participants– and used NGS to identify common mutations. Most mutations, including novel ones, support the idea that those five processes are the ones being affected (Tensen et. al, 2022). Here are the common genes mutated in CTCL (**Table 1**).

Gene Function Groups/Association	Gene Names	Sources
Oncogenes	*TOX, BLK, DEPDC, MYC, PLK1, GTSE1*	Choi et al., 2015; Litvinov et al., 2017; Lai et al., 2021; Motamedi et al., 2021; Park et al., 2021; Tensen et al., 2022; Song et al., 2022; Ren et al., 2023
Tumor Suppressors	*RMB5, TSC1, TP53,*	Choi et al., 2015; Litvinov et al., 2017; Argyropoulos et al., 2020; Motamedi et al., 2021; Park et al., 2021; Tensen et al., 2022; Song et al., 2022; Ren et al., 2023
Chromosome Instability/Remodeling	*BCOR, KDM6A, SMARCB1, TRRAP, DNMT3A, ATM*	Choi et al., 2015; Litvinov et al., 2017; Argyropoulos et al., 2020; Lai et al., 2021; Zhang et al., 2021; Park et al., 2021; Tensen et al., 2022; Song et al., 2022; Ren et al., 2023
Apoptosis	*BIRC5, TRAF1, IFI6, TNFSF11, MELK, BMP2K, HN1, TGFB1/TGFBR1, IL15, STAT1, NPHP3, FAS, PLCG1, TNFAIP3*	Choi et al., 2015; Litvinov et al., 2017; Argyropoulos et al., 2020; Lai et al., 2021; Zhang et al., 2021; Motamedi et al., 2021; Park et al., 2021; Tensen et al., 2022; Song et al., 2022; Ren et al, 2023
NF-κB/JAK-STAT/MAP-kinase signaling	*STAT5B, NFKB2, STAT1, MAP4K3-FIGLA, MAP2K1, NF1, PRKCB, CSNKA1, JAK1, JAK3, STAT3*	Choi et al., 2015; Litvinov et al., 2017; Argyropoulos et al., 2020; Lai et al., 2021; Zhang et al., 2021; Motamedi et al., 2021; Park et al., 2021; Tensen et al., 2022; Song et al., 2022; Ren et al., 2023
Activation Markers	*SELL, CCR6, ITGB1, BRD2, TNFRSF25, RELN, TSPAN2, TNFRSF4, NR4A2*	Litvinov et al., 2017; Argyropoulos et al., 2020; Lai et al., 2021; Zhang et al., 2021; Motamedi et al., 2021; Ren et al., 2023
Immune Markers	*TOX, TIGIT, CTLA-4, PDCD1, LAG3*	Choi et al., 2015; Litvinov et al., 2017; Lai et al., 2021; Motamedi et al., 2021; Park et al., 2021; Tensen et al., 2022; Song et al., 2022; Ren et al., 2023

Table 1. Potential gene mutations in CTCL

However, scientists do not know exactly how everything is connected, especially because there are so many different results, leading to confusion and gaps. There is no clear genetic marker or indication for CTCL and the direct mechanisms of the cancer yet. However, there are many ideas based on the current data given.

Hypothesis: Mutations in critical immune pathways trigger dysregulation of the cell cycle

The typical formation of cancer is when a cell's DNA gets mutated such that the cells go through mitosis much faster. This cell with the mutated DNA then gets selected for in the body, as rapid duplication leads to more chances of survival, invasion, and metastasis. This would eventually build a tumor, and the longer the tumor remains in the body, the more malignant it could become, leading to cancer (Cooper et al., 2020). This suggests that in the formation of cancer, the cell cycle checkpoints is a vital aspect, because the cancer cells bypass them to grow.

My hypothesis states that damage to immune system pathways subsequently affects cell cycle checkpoints, ultimately leading to the development of CTCL. The data suggests that JAK-STAT, NF-κB, and MAP-kinase signaling are affected. JAK-STAT and NF-κB are innate immune systems or the first thing that defend against pathogens, and MAP-kinase signaling determines cell proliferation, differentiation, and death (Alberts et al., 2002; Morrison, 2012). So if these three particular pathways become damaged, antigens might not be fought against and cells could die quickly. Cancer cells as such can also grow rapidly, due to the lack of cell differentiation. This affects the cell cycles and their checkpoints. During the restriction point, apoptosis won't be done to damaged cells, which is why genes like *FAS, PLCG1, TNFAIP3*, and many more are also affected (Manso, 2014; Yenamandra et al., 2022). The data also supports the possibility that affected S-phase genes could result in DNA damage and improper chromatin remodeling. For example, the data suggests that there are mutations in the *DNMT3A* and *ATM* genes. *DNMT3A* plays an integral role in DNA methylation, which determines which gene segments are carried out or silenced (Brunetti et al., 2017). If there is a mutation in this gene, it could result in extra proteins and a lack of necessary proteins, indicating damaged DNA. *ATM* helps with the development of the immune system, damaged DNA detection, and DNA break and repair (Lee et al., 2021). *ATM* mutations could cause damaged DNA to pass through cell checkpoints and inhibit the immune system pathways. Moreover, our body cannot tell apart cells with damaged DNA due to the damaged MAP-kinase signaling pathway.

However, this is simply a hypothesis. We currently don't have clear genes associated with CTCL or indications of the mechanism of the cancer, which leads to a gap. One way to advance the understanding is to use NGS to identify or pinpoint common mutations among a diverse pool of participants. This could lead to a list of common mutations, which leads to a clear diagnosis and genetic markers for the disease. We would also get a better understanding of the cellular processes involved which can inform future scientists and experiments, leading to advanced treatment.

Future Applications of Genetic Testing for CTCL (Discussion)

Due to the complicated amount of gene mutations in CTCL, WGS is the best type of genetic diagnostic tool we can use. Some of these mutations might not be in the exons, which means that exon sequencing isn't specific enough. WGS technology is also getting more and more affordable and common, so this experiment can be performed in a few years. First, we would have to get a decently diverse and large pool of participants who do not have CTCL and compare groups that are more at risk for CTCL than others. Most studies picked out genomes at random, therefore not giving an insight as to why certain groups are more at risk. Picking out specific groups of people who are less and more at risk helps narrow down the many mutations that have been linked with the disease. Given that African American males are at the highest risk for developing CTCL (Markman, 2022), we could compare potential genetic links to CTCL across different genetic ancestries. We would also get different CTCL patients at different stages of the cancer to participate, so that we could see the progression of the cancer, what genes get affected in different stages, and how they show. Most studies take CTCL patients in similar stages, but taking a variety of stages can help show how the cancer cells can accumulate mutations (similar number of people in different stages). We would then compare the genomes to patients at-risk for CTCL and of control patients to identify characteristics or mutations associated with CTCL. This could be accomplished through an analysis like a genome-wide association study (GWAS). We would take samples of the skin, lymph nodes, bone marrow, and the blood, as these are the areas that are most affected by the cancer and are typically biopsied on. Each group would have about fifty people, so that you would see common mutations. This would ensure that other mutations evident in one specific person do not influence the overall data. By comparing common genes from each group, we would be able to figure out which exact genes are being affected in the cancer. This would then lead to clear genetic markers of the cancer, which becomes an effective diagnosis. Clear genetic markers also lead to a better understanding of the mechanisms of the cancer, which then leads to more advanced treatment for earlier stages.

Conclusion

CTCL is an extremely hard cancer to diagnose due to the fact that it presents itself like other skin conditions and cancer cells do not show in the biopsies until late stages, if at all. This creates a huge problem which could be solved with genetic testing. Out of all genetic testing methods, NGS would be a great tool to diagnose CTCL as it can detect exact mutations and it is cost efficient compared to other methods. However, CTCL has multiple genes associated with it, not just one, so it would be hard to find a clear genetic marker. So far, the data suggests that the cell cycle, innate immune systems (JAK-STAT and NFκB), chromatin remodeling, and tumor suppressants are targeted in the cancer. My hypothesis suggested immune system damage and improper cell signaling (MAP-kinase pathway) leads to cell cycle dysregulation. This then leads to cells being unable to perform apoptosis, leading to damaged DNA and cells staying alive in the body, which would ultimately form cancer cells. However, this is just a hypothesis and there is no elucidated mechanism that is associated directly with the cancer. To find a direct correlation

between specific genes and the cancer, I designed an experiment that involved comparing genomes of CTCL patients at different stages of the cancer, “normal” genomes, and genomes that were from more at-risk groups for the cancer. This would show the clear progression of the cancer and lead to a better understanding of CTCL mechanisms.

There are a few limitations to this experiment. The first is the magnitude of this experiment. This is due to the fact that there are 8 specific stages for this cancer (IA - IVB), and then considering the more at-risk and less at-risk groups would lead to around needing at least 600 participants alone for the experiment. Then we would need a big group of scientists to analyze the data, collect samples, etc. Not only is it hard to make an experiment this huge, but it is also hard to get enough people that meet the specific criteria of each group. Another limitation is the cost due to the size of the experiment. WGS technology is typically used for 5-10 genomes. It would be extremely expensive to run 600 genomes through WGS technology. Even though WGS technology costs have decreased dramatically in recent years, it would still cost about $1,000 per test, leading to $600,000 in costs for tests alone (Wetterstrand, 2021). You would also have to get scientists who are willing to work on huge experiments like this, which then again, would become costly very quickly. This would require funding from a government agency, private company, or a research foundation. This designed experiment would be different than what’s currently available for genetic testing, meaning that the patient could be ineligible for insurance coverage for the tests if not for the experiment. Instead, they would be enrolled in this study. If we do find clear indications of the disease through this experiment, then there could be potential FDA-approved DTC tests on the market. Because DTC tests have become very popular in the last decade, having them on the market increases the accessibility of the diagnosis.

CTCL has a survival rate as low as 15% when it has advanced to later stages, but as high as 88% when detected during early stages (Hristov et al., 2023). Developing the process to identify genetic testing markers for CTCL has the power to elucidate new oncogenes, which could impact more than just CTCL patients. Cancer is one of the most elusive, but common diseases in the world. We need to take dramatic, direct steps to understand the development, rapidly diagnose, and improve the treatment of cancer. By using CTCL as a model disease that has been difficult to diagnose, we can apply what we learn to other cancers. Most importantly, developing early detection methods could dramatically improve the lives of the millions of people diagnosed with CTCL each year.

Works Cited

Alberts, Bruce, et al. “Innate Immunity.” *Nih.gov*, Garland Science, 2002, www.ncbi.nlm.nih.gov/books/NBK26846/.

Argyropoulos, Kimon V., et al. “Targeted Genomic Analysis of Cutaneous T Cell Lymphomas Identifies a Subset with Aggressive Clinicopathological Features.” *Blood Cancer Journal*, vol. 10, no. 11, Springer Nature, Nov. 2020, https://doi.org/10.1038/s41408-020-00380-5. Accessed 5 Sept. 2024.

Ataman, Ahmet Dogan, et al. “Medicine in Stamps: History of down Syndrome through Philately.” *Journal of the Turkish German Gynecological Association*, vol. 13, no. 4, Dec. 2012, pp. 267–69, https://doi.org/10.5152/jtgga.2012.43.

Bagherani, Nooshin, and Bruce R Smoller. “An overview of cutaneous T cell lymphomas.” *F1000Research* vol. 5 F1000 Faculty Rev-1882. 28 Jul. 2016, doi:10.12688/f1000research.8829.1

Balasundaram, Palanikumar, and Indirapriya Darshini Avulakunta. “Edward Syndrome.” *PubMed*, StatPearls Publishing, 20 Mar. 2023, www.ncbi.nlm.nih.gov/books/NBK570597/.

Bower, Paul. “Operation of High-Performance Liquid Chromatography (HPLC) | Analytical Chemistry | JoVE.” *Www.jove.com*, 2023, www.jove.com/v/10156/operation-of-high-performance-liquid-chromatography-hplc.

Brlek, Petar, et al. “Implementing Whole Genome Sequencing (WGS) in Clinical Practice: Advantages, Challenges, and Future Perspectives.” *Cells*, vol. 13, no. 6, Jan. 2024, p. 504, https://doi.org/10.3390/cells13060504. Accessed 22 Mar. 2024.

Brunetti, Lorenzo, et al. “DNMT3A in Leukemia.” *Cold Spring Harbor Perspectives in Medicine*, vol. 7, no. 2, Dec. 2016, p. a030320, https://doi.org/10.1101/cshperspect.a030320. Accessed 18 Dec. 2020.

Burnett, David, and John Crocker. *The Science of Laboratory Diagnosis*. Wiley, 2006.

Choi, Jaehyuk, et al. “Genomic Landscape of Cutaneous T Cell Lymphoma.” *Nature Genetics*, vol. 47, no. 9, Nature Portfolio, July 2015, pp. 1011–19, https://doi.org/10.1038/ng.3356. Accessed 30 Oct. 2023.

Cooper, Geoffrey M. “The Development and Causes of Cancer.” *Nih.gov*, Sinauer Associates, 2020, www.ncbi.nlm.nih.gov/books/NBK9963/.

Csanády, László, et al. “STRUCTURE, GATING, and REGULATION of the CFTR ANION CHANNEL.” *Physiological Reviews*, vol. 99, no. 1, Jan. 2019, pp. 707–38, https://doi.org/10.1152/physrev.00007.2018.

Frost, Amy, and Julia van Campen. “Sanger Sequencing — Knowledge Hub.” *GeNotes*, 8 June 2022, www.genomicseducation.hee.nhs.uk/genotes/knowledge-hub/sanger-sequencing/.

Garg, Eshita, and Muhammad Zubair. “Mass Spectrometer.” *PubMed*, StatPearls Publishing, 21 Jan. 2023, www.ncbi.nlm.nih.gov/books/NBK589702/.

George, E., et al. “High Performance Liquid Chromatography (HPLC) as a Screening Tool for Classical Beta-Thalassaemia Trait in Malaysia.” *The Malaysian Journal of Medical*

Sciences : MJMS, vol. 8, no. 2, Penerbit Universiti Sains Malaysia, 2001, pp. 40–46, www.ncbi.nlm.nih.gov/pmc/articles/PMC3413648/.

Griffin, R. Morgan. "Genetic Testing." *WebMD*, edited by Johnson Traci, 9 June 2023, www.webmd.com/baby/genetic-testing.

Hague, Christina, et al. "Cutaneous T-Cell Lymphoma: Diagnosing Subtypes and the Challenges." *British Journal of Hospital Medicine*, vol. 83, no. 4, MA Healthcare, Apr. 2022, pp. 1–7, https://doi.org/10.12968/hmed.2021.0149. Accessed 11 Oct. 2023.

Healy, Marisa. "Cutaneous T-Cell Lymphoma (CTCL): Staging and Treatment | OncoLink." *Oncolink.org*, 29 Apr. 2024, www.oncolink.org/cancers/lymphomas/cutaneous-t-cell-lymphoma-ctcl/cutaneous-t-cell-lymphoma-ctcl-staging-and-treatment. Accessed 5 Sept. 2024.

Heather, James M., and Benjamin Chain. "The Sequence of Sequencers: The History of Sequencing DNA." *Genomics*, vol. 107, no. 1, Jan. 2016, pp. 1–8, https://doi.org/10.1016/j.ygeno.2015.11.003.

Horwitz, Steven. "OBTAINING a PROPER CUTANEOUS LYMPHOMA DIAGNOSIS." *Cutaneous Lymphoma Foundation*, 2019, www.clfoundation.org/obtaining-proper-cutaneous-lymphoma-diagnosis. Accessed 5 Sept. 2024.

Hristov, Alexandra C., et al. "Cutaneous T-Cell Lymphomas: 2023 Update on Diagnosis, Risk-Stratification, and Management." *American Journal of Hematology*, vol. 98, no. 1, Jan. 2023, pp. 193–209, https://doi.org/10.1002/ajh.26760. Accessed 5 Nov. 2023.

Kannan, Thirumulu Ponnuraj, and Bin Alwi Zilfalil. "Cytogenetics: Past, Present and Future." *The Malaysian Journal of Medical Sciences : MJMS*, vol. 16, no. 2, Penerbit Universiti Sains Malaysia, 2009, pp. 4–9, www.ncbi.nlm.nih.gov/pmc/articles/PMC3336168/.

Lai, Pan, and Yang Wang. "Epigenetics of Cutaneous T-Cell Lymphoma: Biomarkers and Therapeutic Potentials." *Cancer Biology and Medicine*, vol. 18, no. 1, 2021, pp. 34–51, https://doi.org/10.20892/j.issn.2095-3941.2020.0216. Accessed 8 May 2022.

Lee, Ji-Hoon, and Tanya T. Paull. "Cellular Functions of the Protein Kinase ATM and Their Relevance to Human Disease." *Nature Reviews Molecular Cell Biology*, vol. 22, no. 12, Dec. 2021, pp. 796–814, https://doi.org/10.1038/s41580-021-00394-2.

Litvinov, Ivan V., et al. "Gene Expression Analysis in Cutaneous T-Cell Lymphomas (CTCL) Highlights Disease Heterogeneity and Potential Diagnostic and Prognostic Indicators." *OncoImmunology*, vol. 6, no. 5, Apr. 2017, p. e1306618, https://doi.org/10.1080/2162402x.2017.1306618. Accessed 16 Nov. 2022.

Ludmann, Paula. "Cutaneous T-Cell Lymphoma: Diagnosis & Treatment." *Www.aad.org*, edited by Aaron Mangold et al., 3 Aug. 2023, www.aad.org/public/diseases/a-z/ctcl-treatment.

Manso, Rebecca. "PLCG1 (Phospholipase C, Gamma 1)." *Atlasgeneticsoncology.org*, 1 Feb. 2014, atlasgeneticsoncology.org/gene/44163/plcg1-(phospholipase-c-gamma-1). Accessed 5 Sept. 2024.

Markman, Maurie. “Cutaneous T-Cell Lymphoma: Mycosis Fungoides & Sezary Syndrome.” *City of Hope*, 12 July 2022, www.cancercenter.com/cancer-types/non-hodgkin-lymphoma/types/cutaneous-t-cell-lymphoma.

Martel, Janelle. “Serum Protein Electrophoresis Test.” *Healthline*, 24 Jan. 2018, www.healthline.com/health/protein-electrophoresis-serum.

Mathew, Mariam T., et al. “Clinical Cytogenetics: Current Practices and Beyond.” *The Journal of Applied Laboratory Medicine*, vol. 9, no. 1, Oxford University Press, Jan. 2024, pp. 61–75, https://doi.org/10.1093/jalm/jfad086. Accessed 16 Feb. 2024.

Moore, Sarah. “What Is Cytogenetics?” *News-Medical.net*, edited by Sophia Coveney, 4 Oct. 2021, www.news-medical.net/life-sciences/What-is-Cytogenetics.aspx.

Morrison, D. K. “MAP Kinase Pathways.” *Cold Spring Harbor Perspectives in Biology*, vol. 4, no. 11, Nov. 2012, pp. a011254–54, https://doi.org/10.1101/cshperspect.a011254.

Motamedi, Melika, et al. “Patterns of Gene Expression in Cutaneous T-Cell Lymphoma: Systematic Review of Transcriptomic Studies in Mycosis Fungoides.” *Cells*, vol. 10, no. 6, Multidisciplinary Digital Publishing Institute, June 2021, pp. 1409–9, https://doi.org/10.3390/cells10061409. Accessed 28 Mar. 2024.

Naeim, Faramarz, et al. *Atlas of Hematopathology*. Editorial: Elsevier, 2013.

Odum, N., et al. “Investigating Heredity in Cutaneous T-Cell Lymphoma in a Unique Cohort of Danish Twins.” *Blood Cancer Journal*, vol. 7, no. 1, Jan. 2017, pp. e517–17, https://doi.org/10.1038/bcj.2016.128. Accessed 25 Apr. 2024.

Park, Joonhee, et al. “Integrated Genomic Analyses of Cutaneous T-Cell Lymphomas Reveal the Molecular Bases for Disease Heterogeneity.” *Blood*, vol. 138, no. 14, Elsevier BV, Oct. 2021, pp. 1225–36, https://doi.org/10.1182/blood.2020009655. Accessed 25 Apr. 2024.

Qin, Dahui. “Next-Generation Sequencing and Its Clinical Application.” *Cancer Biology & Medicine*, vol. 16, no. 1, Chinese Anti-Cancer Association, Feb. 2019, pp. 4–10, https://doi.org/10.20892/j.issn.2095-3941.2018.0055.

Ren, Jingjing, et al. “Integrated Transcriptome and Trajectory Analysis of Cutaneous T-Cell Lymphoma Identifies Putative Precancer Populations.” *Blood Advances*, vol. 7, no. 3, Elsevier BV, Feb. 2023, pp. 445–57, https://doi.org/10.1182/bloodadvances.2022008168. Accessed 5 Sept. 2024.

Retterer, Kyle, et al. “Clinical Application of Whole-Exome Sequencing across Clinical Indications.” *Genetics in Medicine*, vol. 18, no. 7, Dec. 2015, pp. 696–704, https://doi.org/10.1038/gim.2015.148.

Shendure, Jay, and Hanlee Ji. “Next-Generation DNA Sequencing.” *Nature Biotechnology*, vol. 26, no. 10, Oct. 2008, pp. 1135–45, https://doi.org/10.1038/nbt1486.

Shinohara, Michi. “Lab Tests in Staging Cutaneous Lymphoma.” *Cutaneous Lymphoma Foundation*, 2024, www.clfoundation.org/lab-tests. Accessed 5 Sept. 2024.

Song, Xiaofei, et al. “Genomic and Single-Cell Landscape Reveals Novel Drivers and Therapeutic Vulnerabilities of Transformed Cutaneous T-Cell Lymphoma.” *Cancer*

Discovery, vol. 12, no. 5, American Association for Cancer Research, Feb. 2022, pp. 1294–313, https://doi.org/10.1158/2159-8290.cd-21-1207. Accessed 24 Dec. 2023.

Tensen, Cornelis P., et al. "Genetic and Epigenetic Insights into Cutaneous T-Cell Lymphoma." *Blood*, vol. 139, no. 1, Sept. 2021, https://doi.org/10.1182/blood.2019004256. Accessed 14 Nov. 2021.

Wetterstrand, Kris. "The Cost of Sequencing a Human Genome." *Genome.gov*, National Human Genome Research Institute, 1 Nov. 2021, www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost.

Wiese, Daniel, et al. "Disparities in Cutaneous T-Cell Lymphoma Incidence by Race/Ethnicity and Area-Based Socioeconomic Status." *International Journal of Environmental Research and Public Health*, vol. 20, no. 4, Jan. 2023, p. 3578, https://doi.org/10.3390/ijerph20043578. Accessed 26 June 2024.

Winstead, Edward. "FDA Authorizes Blood Test for Hereditary Cancer Risk - NCI." *Www.cancer.gov*, 2 Nov. 2023, www.cancer.gov/news-events/cancer-currents-blog/2023/fda-blood-test-hereditary-cancer-risk.

Zhang, Ping, and Mingzhi Zhang. "Epigenetics in the Pathogenesis and Treatment of Cutaneous T-Cell Lymphoma." *Frontiers in Oncology*, vol. 11, Frontiers Media, June 2021, https://doi.org/10.3389/fonc.2021.663961. Accessed 5 Sept. 2024.

Zimlick, Rachel. "COVID Testing: What's the Difference between Antigens and Antibodies?" *Verywell Health*, 21 June 2024, www.verywellhealth.com/antigen-vs-antibody-7095431.

AI in Healthcare: A Two-Edged Sword By Stanley Huang

Healthcare is the cornerstone of a nation's well-being, safeguarding the fundamental right to life as promised to people in the Declaration of Independence. It functions much like a foundation to a large building, providing support to all structures above it and preventing the deterioration of the building's components. As a nation continues to grow and advance, much like adding new floors and rooms to a building, having a robust healthcare system beneath becomes increasingly important. Currently, however, America is seeing a multitude of cracks in this foundation. Our healthcare system is facing problems such as unequal care, inaccuracies, and lack of efficiency. To combat these problems, people have attempted to integrate Artificial Intelligence (AI) into healthcare, hoping that their new technology will improve the healthcare system. They argue that implementing artificial intelligence in hospitals and clinics will expedite the process of treating patients and increase accuracy of diagnoses, as AI, when built and trained well, can boost efficiency and mitigate errors. Unfortunately, this viewpoint is mistaken. In reality, using artificial intelligence within healthcare can cause more problems, such as exacerbating medical biases, racism, and unequal care, obscuring much-needed transparency to patients and making patient data more susceptible to leaks and hacks. A more measured approach to mending the fractures in our system is needed: AI should act as an assistant to human healthcare providers rather than taking center stage.

As the medical biases and unequal care continue to plague our healthcare system, people have been turning to "fair and unbiased" algorithms in hopes that they may remedy these problems. After all, inanimate machines are free of human bias, right? The truth is, however, that these AI Algorithms unknowingly reflect biases inherent in our current healthcare system because they are fueled and trained by large datasets from patients. A groundbreaking study from 2019 illustrates this discrepancy – a seemingly impartial algorithm built to prioritize patients based on urgency inadvertently propagated medical racism (Obermeyer 447). The algorithm directed less care to Black patients whose conditions were more urgent, because the datasets it was trained with reflected long-standing income and wealth disparities. Other studies, including ones performed on AI designed to detect lung cancer and heart disease, showed that these algorithms delayed life-saving treatment to minorities, especially women and people of color, since a vast majority of the datasets used were taken from Caucasian males (Larrazabal 12592). These studies and many others all suggest that allowing artificial intelligence too much control over healthcare is unethical, because algorithms, most of which are trained on biased datasets, will exacerbate the bias and discrimination within the healthcare system.

It may then seem logical to use "fairness algorithms", or in other words, alter the algorithm's detection sensitivity for disadvantaged or advantaged groups. By adjusting algorithm sensitivity, one might think, healthcare AI developers can mitigate bias against disadvantaged groups, while still using the AI to diagnose patients. However, these fairness algorithms actually cause more problems. There are multiple ways that fairness algorithms can work: they may either increase performance for historically disadvantaged groups or they may decrease performance for advantaged groups. Both are problematic. The former, which involves altering the results of the disadvantaged groups to be

more sensitive in identifying diseases, comes at the cost of accuracy. Let's say we have a cancer-detection algorithm which has higher accuracy when observing white patients as opposed to black patients. In order to counteract this bias, which may be caused by unrepresentative or inaccurate data sets, the algorithm may try to err on the side of caution and diagnose more black patients with cancer. However, this algorithm increases the chance for false positives, events in which the algorithm detects a disease which the patient was never afflicted with. Indeed, in a study co-authored by Harvard researchers, using fairness algorithms resulted in worse model performance across several healthcare algorithms (Chen 719).

The other option, to decrease the algorithm sensitivity for advantaged groups, is even more detrimental, as it may deny life-saving treatment to patients. In other words, this attempt to balance the scales not only comes at the cost of human life, but also ignores the principle which our founding fathers deemed important: that everyone deserves a right to life and by extension, access to healthcare when needed.

It seems, therefore, that to implement artificial intelligence within healthcare creates a dilemma: we either accentuate medical bias, or we sacrifice precision in a field where lack of precision means death. By using fairness algorithms to remedy the inherent bias that AI algorithms exhibit, one is essentially attempting to orally remove venom from a snakebite: in trying to remove a poison deeply entrenched within the system, they inadvertently increase risk of harm or even death. Of course, the most sensible option is to avoid meddling with this snake in the first place — to avoid relying on artificial intelligence in our healthcare system altogether.

Perhaps one of the main obstacles to tackling the issue of medical bias in artificial intelligence is that AI lacks transparency and is difficult to understand. In fact, many of the companies that develop AI related healthcare products work in a "black box," or a complex piece of equipment which has components that are mysterious to the user. Without access to the datasets which the AI was trained on, or understanding how the algorithm makes its conclusions, how can one go about identifying and mitigating flaws? These black boxes are more detrimental than they seem: the lack of transparency raises the risks of a patient incorrectly or inappropriately using these products which may lead to improper care and inaccurate diagnoses. In an investigation of 118 FDA-approved AI healthcare products, only nine provided basic racial demographic data to prove the validity of that product (Ebrahimian 560). This oversight by the FDA could prove fatal in cases where a disease, such as breast cancer (which is more dangerous for black women (Siddarth 514)), has a disparate impact on underrepresented groups. But simply trying to explain these "black box" algorithms is insufficient: "trying to explain black box models, rather than creating models that are interpretable in the first place, is likely to perpetuate bad practice and can potentially cause great harm to society" (Rudin 206). In other words, instead of attempting to understand these cryptic and obscure systems, why not simply build and use systems which we understand and are familiar with? Before we should trust AI algorithms in our healthcare system, they must become more transparent to both prescribers and patients alike.

To navigate the treacherous uncharted waters of AI implementation within our healthcare system, we must take care when placing precautions on these algorithms. If a clinician wanted to

test a patient for sepsis, a critical, life-threatening complication, using an artificial intelligence algorithm, the prescriber and patient must know how the developers tested and evaluated their algorithm, as well as how well it performed. However, in our current healthcare system, no such North Star exists to guide our journey. Instead, dark clouds block our vision, putting the patient at risk of harm or even death. Take the infamous Epic Sepsis Model, for example, which researchers at a hospital chain found to only succeed in detecting sepsis in infected patients a mere one-third of the time (Wong 1065). In other words, if this model tests 100 patients for sepsis, roughly 66 of them would test negative and continue living without the knowledge that they are afflicted with a condition that causes organ failure and therefore requires immediate attention. In addition, AI developers must also state how patients should use their algorithms, especially for important information such as the intended patients or the conditions under which the developers tested the algorithm. Not doing so leaves the life of the patient at the whim of incomprehensible machinery. Clinicians should also be able to access the data used to train the AI, which could allow them to determine the tools that certain patients need. This transparency could help mitigate the disparities in the healthcare system, as physicians could make sure that any given algorithm will work on a particular patient. Without transparency or accountability, relying on artificial intelligence in healthcare is akin to navigating a treacherous ocean blindfolded and without a compass, putting the sailors on board at risk of severe danger.

In addition to perpetuating medical biases and putting patients at risk of improperly using these algorithms, using AI in healthcare puts patient data at immense risk. Data breaches occur within the US healthcare system very often, to the point that people have created a whole industry to defend against cybercriminals attempting to steal patient data. After all, patient data is perhaps the most valuable form of information nowadays. As Tom Kellerman, chief cybersecurity officer of VM Carbon Black, a company that develops security software, puts it: “Health information is a treasure trove for criminals. By compromising it, by stealing it, by having it sold, you have seven to 10 personal identifying characteristics of an individual” (Steger). Kellerman’s claim may sound bold, but with stolen healthcare data, hackers can commit tax fraud, home loan equity fraud and even identity theft, not to mention financially extorting the hacked individual. This industry is so lucrative, in fact, that healthcare data sells on the black market for more than social security and credit card numbers. In 2022 alone, 707 major healthcare-related data breaches occurred, affecting 49.6 million Americans, according to the US Department of Health and Human Services’s Office for Civil Rights. In addition, the number of major breaches has sharply increased in the last decade and a half (with 2009 having a mere 18 reported breaches) (Murray-Watson). Similarly, in this time frame, hospitals and clinics have begun to rely on AI more. This correlation — between increase of healthcare data breaches and stronger reliance on AI — suggests how dangerous the use of AI could become in the future.

How does relying on AI actually lead to immense data vulnerabilities? In order to train these algorithms, developers must pass vast amounts of patient data, which they compile in health records or cloud-based systems that become lucrative targets for cybercriminals. Not only does the use of these algorithms collect patient data for cybercriminals to take, but using transparency-lacking

algorithms can cause problems when those handling that intelligence are unaware of how to control it or are unaware of its risk. Reliance on AI algorithms in healthcare compromises the safety of patient data in other ways as well. Kellerman claims that “The big challenge with the entire governance of the healthcare sector with regards to cybersecurity, is that ... [those who run the board] are very astute when it comes to medical knowledge but not quite prepared to handle the risks of IT and IT deployment.” (Steger). In this sense, using artificial intelligence is like trying to hide valuable possessions from thieves by placing them all in one safe, with little or no protection, whose location is openly broadcast.

So if reliance on artificial intelligence within the healthcare system is dangerous, is there any way to use it well? Surely such a useful advancement can offer some assistance in the healthcare field without too many drawbacks. As it turns out, while using AI extensively in healthcare to make important decisions and diagnose patients can be risky, employing AI as an assistant to support healthcare professionals can enhance their capabilities. Since algorithms are far more capable than humans at storing information and even analyzing data, using AI to influence and support the decisions of healthcare professionals takes advantage of this benefit while mitigating possible risks. Rather than letting the algorithms diagnose the patients, doctors should diagnose the patients with the help of algorithms that can generate plausible diagnoses for the doctor to approve. In fact, Oracle, a company best known for their computer software, has recently created an algorithm that performs this exact task. When prompted, the Oracle AI automates note-taking, and when prompted, looks up required elements in its database and returns necessary information, giving suggestions which will allow the physicians to better craft a treatment plan (Landi). In this scenario, healthcare professionals are using algorithms to their fullest ability without losing the benefits that an in-person healthcare professional can provide, which will increase accuracy of diagnosis and avoid the dangers of letting AI control the process. In other words, let the algorithms be an instrument in the orchestra, not the conductor.

Artificial intelligence is a rapidly growing tool with countless applications in which it has proved to be of considerable benefit, but those who believe that using artificial intelligence within the healthcare system will solve the system’s problems are sorely mistaken. These supporters may think implementing AI in healthcare is akin to applying penicillin to kill bacteria, the use of a powerful and novel medication to eliminate weaknesses and problems. Instead, AI is more like a shoddy bandage, which simply covers the wound rather than treating it, and can lead to infection and other complications when used too extensively. Even so, a shoddy bandage does have its uses. After being sterilized and after antiseptics and antibiotics are applied along with it, it can treat wounds, or in this case, help save lives. As we continue to develop and improve our technologies, we must remember one core principle: AI, like any tool, is imperfect. But by using AI as an assistant, we acknowledge its potential while also preventing the numerous negative side effects of relying on it in healthcare.

Works Cited

Chen, Richard, et al. “Algorithmic Fairness in Artificial Intelligence for Medicine and Healthcare.” Nature Biomedical Engineering, vol. 7, 2023, pp. 719–742, https://doi.org/10.1038/s41551-023-01056-8. Accessed 6 Dec. 2023.

Ebrahimian, Shadi, et al. “FDA-Regulated AI Algorithms: Trends, Strengths, and Gaps of Validation Studies.” Academic Radiology, vol. 4, 2021, pp. 559–566, https://doi.org/10.1016/j.acra.2021.09.002. Accessed 13 Dec. 2023.

Landi, Heather. “Oracle Health Integrates Generative AI, Voice Tech into EHR System to Automate Medical Note-Taking.” Fierce Healthcare, 20 Sept. 2023, https://www.fiercehealthcare.com/ai-and-machine-learning/oracle-health-integrates-generative-ai-conversational-voice-tech-ehr-system. Accessed 13 Dec. 2023.

Larrazabal, Agostina, et al. “Gender Imbalance in Medical Imaging Datasets Produces Biased Classifiers for Computer-Aided Diagnosis.” Proceedings of the National Academy of Sciences, vol. 117, no. 23, 2020, pp. 12592–12594, https://doi.org/10.1073/pnas.1919012117. Accessed 6 Dec. 2023.

Murray-Watson, Rebecca. “Healthcare Data Breach Statistics.” The HIPAA Journal, 23 Nov. 2023, https://www.hipaajournal.com/healthcare-data-breach-statistics. Accessed 13 Dec. 2023.

Obermeyer, Ziad, et al. “Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations.” Science, vol. 366, no. 6464, 2019, pp. 447–453, https://doi.org/10.1126/science.aax2342. Accessed 6 Dec. 2023.

Rudin, Cynthia. “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.” Nature Machine Intelligence, vol. 1, 2019, pp. 206–215, https://doi.org/10.1038/s42256-019-0048-x. Accessed 13 Dec. 2023.

Siddarth, Sumit, et al. “Racial Disparity and Triple-Negative Breast Cancer in African-American Women: A Multifaceted Affair between Obesity, Biology, and Socioeconomic Determinants.” Cancers, vol. 10, no. 12, 2018, p. 514, https://doi.org/10.3390/cancers10120514. Accessed 13 Dec. 2023.

Steger, Andrew. "What Happens to Stolen Healthcare Data?” HealthTech, 30 Oct. 2019, https://healthtechmagazine.net/article/2019/10/what-happens-stolen-healthcare-data-perfcon. Accessed 13 Dec. 2023.

Wong, Andrew, et al. “External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients.” JAMA Internal Medicine, vol. 181, no. 8, 2021, pp. 1065–1070, https://doi.org/10.1001/jamainternmed.2021.2626. Accessed 13 Dec. 2023.

Biomechanical Factors Influencing Speed and Accuracy of Water Polo Shots
By William Schoinas

Abstract

Water polo shots require the coordination of muscle groups across the entire body. However, the specific contributions of these muscles to shot speed and accuracy remain poorly understood. This review synthesizes findings from electromyographic analyses, motion studies, and anthropometric research to explore the biomechanical factors underlying water polo shots. Electromyographic studies provide insight into patterns of muscle activation, while motion analyses focus on the role of upper-body kinematics. Anthropometric research highlights the importance of body strength and coordination. However, conflicting evidence about the roles of muscles like the rectus abdominis and rectus femoris underscores the need for deeper investigation. By integrating current findings, this review identifies gaps in the literature and offers actionable insights to optimize water polo performance. Future research should focus on advanced techniques such as underwater EMG and integrated motion-muscle studies to refine our understanding of the mechanics that drive elite water polo performance.

Introduction

Water polo shots involve a transfer of power through torque using many muscles throughout the entire body, not just the shoulder and arm (Jayanti et al.; Solum). However, neither Jayanti et al. nor Solum state the contribution of different muscle groups to shooting performance. As a competitive water polo player, I often wondered whether science can provide the answer. Hence, this literature review aims to determine the primary muscle groups that are predominantly involved in influencing the accuracy and speed of a water polo shot in order to better understand the water polo shot biomechanics and thereby, allow players to gain a competitive advantage.

This topic and research question are worthy of investigation due to a general lack of understanding of the biomechanics involved in a water polo shot throughout the water polo community. A better understanding of these biomechanics could help players gain a competitive advantage in their sport.

Water Polo Shot Biomechanics

Water polo experts describe the overhand water polo shot as a movement with complex biomechanics that involves the whole body (Hsu et al.; Jayanti et al.; Solum). According to Dr. Solum, the overhand water polo shot is split into three phases. Moreover, he comprehensively lists the sequence of lower and upper movements involved in a typical water polo power shot during these phases (Solum).

The phases of a shot are as follows (Solum): the **elevation phase**, where a player uses the eggbeater kick to elevate out of the water, the **rotation phase**, where a player uses breaststroke kicks to elevate quickly and rotates their body to generate torque, and finally the **shooting phase**,

where the player is still maintaining their height with an eggbeater kick and pushes and releases the ball forward finishing with their arm in the water.

Dr. Jim Solum comprehensively lists the sequence of lower and upper movements involved in a typical water polo power shot during these phases (Solum). With regards to the lower body movements, "the shooter points the left foot at the goal, the right leg is straight back, the right foot rotates inward and outward to kick, the hips rotate back to cock the body and rotate forward to throw the ball. The weight of the body transfers from the right foot to cock the ball to the left foot to help release the ball" (Solum).

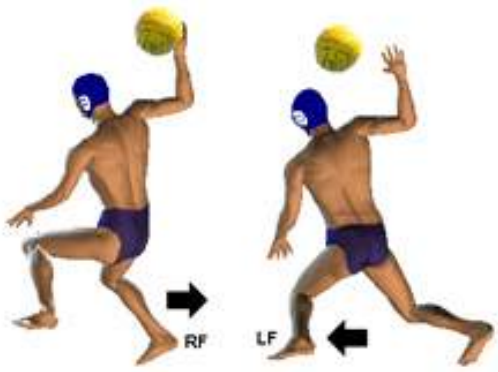


Figure 1: Lower Body Movements (Solum)

With regards to the upper body movements, "the shooter's five upper body fundamentals position the back in the vertical, the abs crunch and contract to snap the torso forward, the left hand sweeps to the left to turn the body to the right to cock the right arm and then pulls down to elevate and rotate the hips. The shooter's right arm is high in the air, close to the ear, with the elbow at ear level. The right hand grips the ball softly in a horizontal cradle position or in a vertical hand position with the finger firmly pinching the ball. The ball is released using the standard 3-finger release, a 2-finger release or 1-finger release" (Solum).

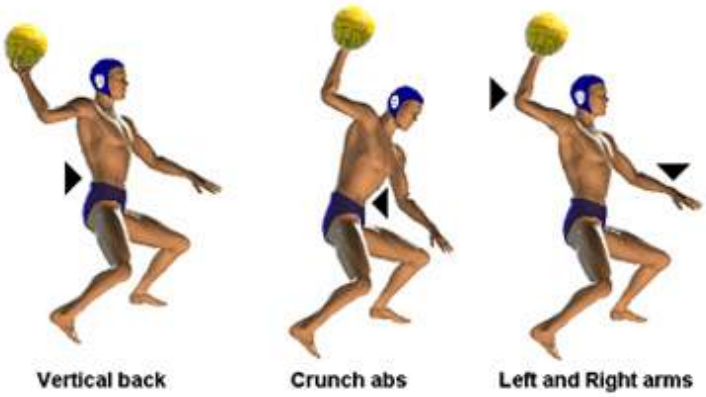


Figure 2: Upper Body Movements (Solum)

A training rubric developed through interviews of water polo coaches reinforces the contribution of the full body to the shot performance by drawing attention to coaches to the placement and movement of all body parts during practices (Jayanti et al.). The authors mention that the contribution of the torso to the shot speed is about 30% (Jayanti et al.). However, no scientific studies are cited to support this claim.

Clearly, water polo practitioners focus on the biomechanics of the entire body in optimizing shot performance. Academic research in water polo shots can be classified into three categories: (a) **muscle activity research** uses electromyographic (EMG) measurements to determine the activity level of different muscle groups during water polo shots, (b) **motion analysis research** uses high-speed photography to determine the motion characteristics of different body parts and the water polo ball during water polo shots, and (c) **body and physiological measurements research** examine the correlation between body measurements (e.g., arm girth) and physiological measurements (swimming speed) with water polo shot performance.

Muscle Activity

Clarys et al. published one of the first studies analyzing the speed and accuracy of overhand water polo shots (Clarys et al.). The study participants consisted of ten elite male water polo players from the A1 and A2 divisions in the Belgian league. The subjects were asked to throw the ball from 4m and 8m away towards a force plate target in the goal area. Surface EMG measurements were used to detect the activity intensity of the arm, trunk, and lower limb muscles during the throw. Precision was approximated from the plate impact force. At 4m, positive correlations were reported between impact force and biceps brachii or triceps brachii (upper arm) muscle intensity, and negative correlations were reported with rectus abdominis (torso) muscle intensity. At 8m, positive correlations were reported between impact force and biceps brachii or pectoralis major (shoulder) muscle intensity, and negative correlations were reported with rectus abdominis and rectus femoris (lower leg) muscle intensity (Clarys et al.).

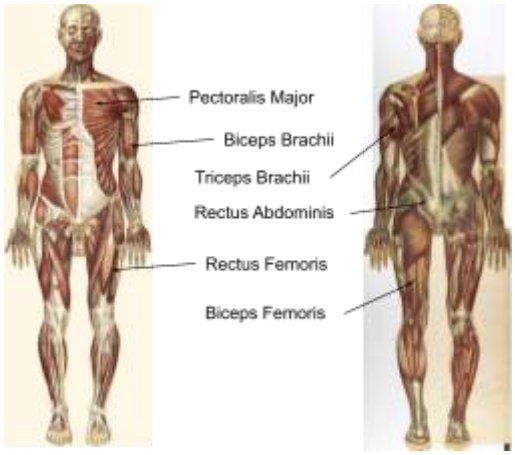


Figure 3: Major Muscle Groups measured in (Clarys et al.).
Images adapted from (*Historical Anatomies on the Web*).

The authors attributed the correlation differences at the two distances to the more significant effort required to hit the target at longer distances. The authors also acknowledged that their negative correlation results directly contradict the beliefs of coaches and players. This surprising result has not been further explored in academic literature, and in almost three decades, there are no indications that it has had any effect on water polo training practices (Clarys et al.).

Muscle	**Impact force**		**Precision**	
	4m	**8m**	**4m**	**8m**
Rectus femoris	ns	-0.56	-0.89	ns
Biceps femoris	ns	ns	0.94	ns
Rectus abdominis	-0.99	-0.96	-0.99	-0.99
Pectoralis major	ns	0.87	ns	0.67
Biceps brachii	0.76	0.73	ns	0.58
Triceps brachii	0.58	ns	0.71	0.63

Table 1: Muscle intensity (EMG) correlations ($p \leq 0.05$) with impact force and precision scores. Adapted from (Clarys et al.).

A more recent paper measured electromyographic parameters such as the percentage of time that the muscle is activated during the shot and the normalized electrical activity (EMG amplitude) of the muscle, expressed as a percentage of its maximum voluntary contraction (MVC) for selected shoulder and arm muscles during different types of water polo shots. (Yaghoubi, Moghadam, et al.)

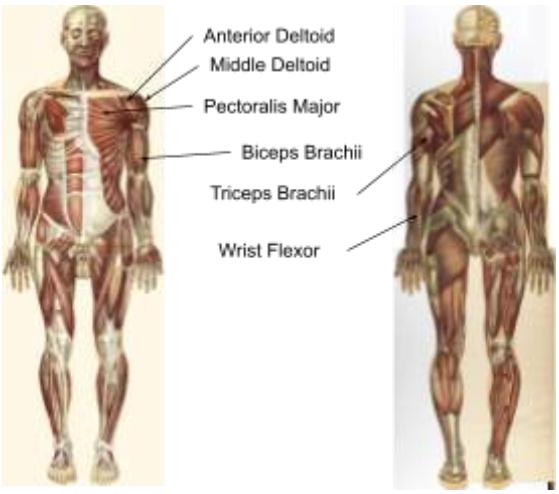


Figure 4: Major Muscle Groups measured in (Yaghoubi, Moghadam, et al.).
Images adapted from (Historical Anatomies on the Web).

The study subjects were 12 experienced players from the Iranian national team. While the data reveal many details about the muscle contractions during the shots, the measurements were not correlated with shot performance metrics(Yaghoubi, Moghadam, et al.). It is also difficult to compare these results to the muscle intensity metric in Clarys et al. because different muscle groups were measured. In contrast to the Clarys et al. paper, the authors note the contribution of the entire body in short performance and suggest this is an area for further research (Yaghoubi, Moghadam, et al.).

Muscle Groups	**Pectoralis Major**		**Anterior Deltoid**		**Middle Deltoid**	
Water Polo Shots	**Duration (%)**	**Amplitude (%)**	**Duration (%)**	**Amplitude (%)**	**Duration (%)**	**Amplitude (%)**
Penalty shot	47.97 (18.56)	37.16 (8.26)	70.96 (25.52)	22.44 (5.47)	65.83 (17.87)	14.25 (6.17)
Overhead shot	61.02 (21.30)	43.60 (14.49)	62.73 (21.24)	26.30 (7.89)	58.52 (11.10)	16.31 (6.53)
Push shot	44.76 (7.46)	42.23 (18.15)	78.98 (19.86)	33.60 (9.90)	78.87 (11.65)	17.60 (6.11)
Back shot	35.08 (25.26)	20.95 (12.58)	45.00 (23.78)	24.78 (6.10)	79.46 (20.58)	16.77 (5.79)

Table 2a: Mean (SD) of duration and average normalized values of the muscles examined in different water polo shots. Adapted from (Yaghoubi, Moghadam, et al.).

Muscle Groups	**Biceps Brachii**		**Triceps Brachii**		**Wrist Flexor**	
Water Polo Shots	**Duration (%)**	**Amplitude (%)**	**Duration (%)**	**Amplitude (%)**	**Duration (%)**	**Amplitude (%)**
Penalty shot	73.65 (24.96)	16.43 (9.81)	46.41 (12.24)	38.78 (24.91)	62.61 (24.47)	24.99 (10.63)
Overhead shot	49.63 (18.97)	24.65 (11.20)	52.20 (14.83)	35.64 (20.12)	66.09 (13.00)	37.08 (16.59)
Push shot	66.28 (6.89)	16.09 (11.64)	75.38 (15.59)	43.38 (25.64)	62.52 (17.05)	38.05 (11.34)
Back shot	54.12 (24.00)	21.93 (11.82)	73.84 (17.61)	36.90 (23.64)	59.96 (22.42)	54.48 (25.65)

Table 2b: Mean (SD) of duration and average normalized values of the muscles examined in different water polo shots. Adapted from (Yaghoubi, Moghadam, et al.).

A subsequent paper contrasted the electrical activities of key muscle groups between twelve experienced water players and ten non-players (Yaghoubi, Esfehani, et al.). The authors observed that the "normalized electrical activities of triceps brachii, pectoralis major, and wrist flexors were greater than other muscles" in experienced water polo players. In contrast, only the triceps brachii played an important role in the control group. More importantly, the authors note that "only experienced water polo players activated the observed muscles in a specific sequence, from proximal to distal," (Yaghoubi, Esfehani, et al.)**.**

	Time to Peak					
Muscle Group	Pectoralis Major	Anterior Deltoid	Middle Deltoid	Biceps	Triceps	Wrist Flexors
Inexperienced	0.30 (0.17)	0.26 (0.19)	0.15 (0.07)	0.15 (0.13)	0.23 (0.18)	0.31 (0.19)
Expert	0.19 (0.09)	0.10 (0.06)	0.08 (0.04)	0.12 (0.10)	0.08 (0.04)	0.17 (0.08)

Table 3a: Mean (SD) of the time to reach the maximum activation (time to peak). Adapted from (Yaghoubi, Esfehani, et al.)

	Time Broadness %	**MVC %**					
Muscle Group		Pectoralis Major	Anterior Deltoid	Middle Deltoid	Biceps	Triceps	Wrist Flexors
Inexperienced	26.31 (11.40)	37.76 (7.69)	41.04 (6.92)	28.43 (6.30)	42.56 (12.56)	68.51 (18.58)	35.56 (7.60)
Expert	18.84 (8.69)	43.60 (14.49)	26.30 (7.89)	16.31 (6.53)	24.65 (11.20)	35.64 (20.12)	37.08 (16.59)

Table 3b: The percentage of the total movement time during which a muscle remains close to its peak activation level (time broadness) and the muscle's activity level during the movement, expressed as its MVC percentage. Adapted from (Yaghoubi, Esfehani, et al.)

Motion Analysis

Feltner and Taylor analyzed the upper arm (shoulder, elbow, wrist) resultant joint forces and resultant joint torques in penalty shots across two shooting techniques, the overhand power shot, and the sweep shot (Feltner and Taylor). The study relied on high-speed photography and visual markers affixed on the upper body of thirteen intercollegiate water polo players to determine the motion of the upper body during a water polo shot. The choice of the shot technique was left up to the individual players. The level of player proficiency was not disclosed in the paper, but the average time spent playing water polo was cited at around six years, indicating relatively inexperienced players compared to elite players in competitive league teams (Feltner and Taylor).

The authors reported that the percentage contribution of upper arm internal/external rotation angular velocity to the speed of the ball at release is an indicator of shot technique and that there is a strong positive correlation between strength indicators (circumference of chest, axillary, biceps, elbow, forearm wrist) or physical characteristics (forearm and middle finger length) with shot technique. The authors explained that strong players are able to use the overhand technique to throw the ball at sufficient speed, while less strong players have to rely on the sweep technique to do so (Feltner and Taylor).

Measure	Mean	SD	Correlation (r)
Mass (kg)	81.3	9.5	0.50
Height (cm)	185.4	7.0	0.36
Circumference Measures (cm)			
- Waist	85.2	6.4	0.34
- Chest	100.1	5.7	0.69 *
- Axillary	33.4	2.2	0.57 *
- Biceps	31.7	1.7	0.58 *
- Elbow	27.9	1.3	0.62 *
- Forearm	27.8	1.2	0.55 *
- Wrist	17.5	0.7	0.63 *
- Hand	21.6	0.8	0.45
Segment Lengths (cm)			
- Upper Arm	35.2	2.8	0.05
- Forearm	27.8	1.7	0.59 *
- Hand	8.8	0.5	0.07
- Finger III (Middle Finger)	11.2	0.5	0.72 *
Hand Breadth (cm)	8.7	0.4	0.22
Hand Span (cm)	22.8	1.2	0.16

Table 4: Anthropometric Data and their Correlations with the Percentage Contribution of Upper Arm Internal Rotation $\omega_{UA:IE}$ to Ball Speed $|v_B|$. Adapted from (Feltner and Taylor).
($* p \leq 0.05$)

Papers published in subsequent years used high-speed photography to analyze the movement of the upper body during water polo shots. Hsu et al. analyzed the movement of upper extremities using high-speed cameras (Hsu et al.). The paper focused on the physical characteristics of the movement in terms of speed and angle but had no further insights into biomechanics. Van den Tillaar contrasted the movement of different body parts during throwing across different sports. For water polo, the paper assumed a simplified model of overhead shots involving only the upper extremities rather than the power shorts typically used in competitive games involving the whole body. Accordingly, it drew no conclusions applicable to competitive water polo (Van den Tillaar).

Melchiorri et al. compared the physical motion during water polo shots between elite and sub-elite players and concluded that the difference in shot performance can be explained by contrasting the throwing kinematics (e.g., body weight had no impact) (Melchiorri et al.). Differences in kinematics between the two groups were reported. Specifically, elite players had "a greater angle at release, shorter throwing time, smaller shoulder loading angle, and a greater head height at the time of the shot." The authors attributed the difference in shot performance to the greater head height and greater elbow angle at the time of ball release (Melchiorri et al.).

Body and Physiological Measurements

McCluskey et al. focused on female players and correlated shot speed to physiological characteristics as well as land jump height (McCluskey et al.). Subjects were recruited from participants in the Western Australian Water Polo Incorporated A Grade summer competition, Western Australian Institute of Sport water polo scholarship holders, and members of the Western Australian Under-17 water polo team. The authors found a significant correlation between shot speed and height, weight, relaxed arm girth, calf girth, and land jump height. This is a clear indication that the lower extremities and body core have a significant contribution to shot performance (McCluskey et al.).

	Whole group		**Throwing velocity >15.3 m/s**		**Throwing velocity ≤15.3 m/s**		**P values**
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Age (year)	20.4	6.16	20.73	6.36	20.09	6.25	0.82
Height (cm)	170.7	6.23	173.88	4.99	167.61	5.91	0.01 *
Weight (kg)	68.3	8.87	73.11	9.32	63.45	5.19	0.01 *
Upper limb length (cm)	75.6	3.42	76.52	3.76	74.70	2.95	0.22
Lower limb length (cm)	106.2	6.42	108.14	6.18	104.28	6.33	0.16

Sitting height (cm)	90.2	4.20	91.85	2.18	88.47	5.09	0.06
Fat mass (kg)	13.8	3.66	14.9	3.12	12.6	3.56	0.12
Lean mass (kg)	54.5	6.57	58.18	6.05	50.68	3.81	<0.001 *
Chest girth (cm)	94.1	5.09	95.85	5.33	91.94	4.06	0.09
Waist girth (cm)	73.4	5.72	75.18	6.88	71.30	3.03	0.13
Gluteal girth (cm)	96.9	9.21	99.27	9.70	94.09	8.20	0.22
Arm (relaxed) girth (cm)	30.4	2.42	31.49	2.37	28.97	1.73	0.02 *
Thigh gluteal girth (cm)	58.9	4.66	60.42	4.87	56.96	3.81	0.10
Calf girth (cm)	35.6	2.52	36.81	1.95	34.10	2.40	0.01 *
Biacromial breadth (cm)	37.7	1.74	38.33	1.83	36.98	1.37	0.08
Transverse chest breadth (cm)	27.8	2.61	28.45	3.01	27.07	1.90	0.25
Biiliocristal breadth (cm)	28.2	2.01	28.44	2.21	27.87	1.82	0.54
Femur breadth (cm)	9.5	0.64	9.63	0.65	9.34	0.63	0.34
Humerus breadth (cm)	6.6	0.44	6.74	0.42	6.37	0.38	0.06
Land jump height (cm)	37.95	5.1	40.82	3.97	35.09	4.48	<0.001 *
Power (kW)	3.342	0.558	3.735	0.504	2.949	0.245	<0.001 *

Table 5: Characteristics of Participants According to Throwing Velocity. Adapted from (McCluskey et al.). (* $p \leq 0.05$)

Platanou and Varamenti focused on female players and correlated physiological and technical characteristics to shot performance (Platanou and Varamenti). The subject set consisted of 33 elite-level members of the first four teams of the Greek A1 women's league. The authors found a positive correlation between body length as well as swimming speed, internal and external shoulder torque, and VO_2. A positive correlation was also found with the length of the

hand. Lastly, a positive correlation was found between the ball's speed and the length of the biacromial breadth. The paper speculates that this is due to the fact that the size of the biacromial breadth is also related to the shoulder's internal and external torque force (Platanou and Varamenti).

Variables	**Throwing Velocity**
Speed 25 ($m \cdot s^{-1}$)	0.42
Internal Rotation (N·m)	0.70
External Rotation (N·m)	0.62
VO_2 ($l \cdot min^{-1}$)	0.56
Lower Limb	0.37
Hand	0.41
Biacromial Breadth	0.44

Table 6: Pearson's Statistically Significant Correlation Coefficients between Physiological and Anthropometric Characteristics with Throwing Velocity. Adapted from (Platanou and Varamenti).

Ferragut Fiol et al. correlated physiological measurements with shot performance in different shots (shots with and without a goalkeeper and dynamic shots) for male players playing in different water polo positions (goalkeeper, center back, center forward, wing) (Ferragut Fiol et al.). The subject set consisted of 94 injury-free male water polo players who were playing in the Spanish King´s Cup and were grouped as follows: 15 goalkeepers, 45 offensive wings, 20 center backs, and 14 center forwards. A significant correlation was found for center backs with flexed and tensed arm girth and hand grip and for wings with height, body mass, muscular mass, relaxed arm girth, flexed and tensed arm girth, forearm girth, wrist girth, chest girth, bi-acromial breadth, biiliocristal breadth. No significant correlations were found for center forwards or goalkeepers. While the authors did not provide an explanation for the difference in results between player positions, goalkeepers, center forwards, and center backs have specialized roles that do not depend on mastering water polo shots perfectly. Focusing on offensive wing players, muscular mass, arm relaxed girth, and arm flexed and tensed girth correlate the strongest with shot performance, especially in dynamic shots, which correspond to overhand power shots in other studies (Ferragut Fiol et al.).

Player Position	**Anthropometric Characteristics**	**Throwing without Goalkeeper**	**Throwing with Goalkeeper**	**Dynamic Shot**
Center Back (n=17)	**Arm girth flexed and tensed girth**	0.522*	ns	ns

	Hand Grip	0.501*	ns	ns
Wing (n=39)	**Height**	ns	0.364*	0.450*
	Body Mass	0.389*	0.441*	0.493*
	Arm Span	0.326*	0.382*	0.335*
	Iliac Crest skinfold	0.321*	0.337*	ns
	Abdominal skinfold	0.321*	0.372*	0.339*
	Muscular Mass	ns	0.468*	0.504*
	Arm relaxed girth	0.361*	0.433*	0.534*
	Arm flexed and tensed girth	ns	0.383*	0.510*
	Forearm girth	ns	0.367*	0.476*
	Wrist girth	0.394*	0.372*	0.442*
	Chest girth	ns	ns	0.427*
	Waist girth	ns	0.317*	ns
	Biacromial breadth	0.462*	ns	ns
	Chest AP breadth	0.390*	0.352*	0.386*
	Biliocristal breadth	ns	ns	ns
	Arm length	ns	0.384*	0.358*
	Hand grip	ns	0.355*	0.353*

Table 7: Correlation coefficient values obtained between anthropometric variables and throwing speed in the three tested conditions. Adapted from (Ferragut Fiol et al.). ($* \ p \leq \ 0.05$)

Discussion & Conclusions

Research measuring muscle activity during water polo shots is relatively rare. Only (Clarys et al.) attempted to collect measurements from all major muscle groups, and their results on the negative correlation between rectus abdominis and rectus femoris activity with shot performance were not replicated in later studies. Yaghoubi, Esfehani, et al.; Yaghoubi, Moghadam, et al. did not correlate EMG measurements with shot performance, and they were restricted to upper body measurements. Nevertheless, (Yaghoubi, Esfehani, et al.) offers the first significant insight by comparing the muscle activation sequence between experienced water polo players and non-players. The authors state that experienced water polo players activate more

muscle groups among the upper body muscle groups than non-players, and the sequence of activation is from distal to proximal (Yaghoubi, Esfehani, et al.).

Research using motion analysis to understand the biomechanics of water polo shots is more common than research measuring muscle activity. However, all research is focused on upper-body motion that is visible without an underwater camera. McCluskey et al. correlate motion measurements with the use of overhead or sweeping shot techniques (McCluskey et al.). While interesting, the results are not useful for this review since the contribution of different muscle groups should be different in different shot techniques. Melchiorri et al. contrasted elite and sub-elite players noting that elite players had "a greater angle at release, shorter throwing time, smaller shoulder loading angle, and a greater head height at the time of the shot" (Melchiorri et al.). This result is consistent with Yaghoubi, Esfehani, et al.'s results because in order to achieve a greater angle at release, shorter throwing time, and greater head height at the time of the shot, more muscle groups must be activated (Yaghoubi, Esfehani, et al.) .

Anthropometric and physiological research is also common. McCluskey et al. found a significant correlation between shot speed and height, weight, relaxed arm girth, calf girth, and land jump height (McCluskey et al.). This indirectly means that the torso and lower extremities contribute to shot performance since presumably bigger muscles can exert high forces. Therefore, this result contradicts Clarys et al.'s results about the negative impact of higher muscle activity intensity by the same muscle groups on shot performance (Clarys et al.). Platanou and Varamenti also focus on female players and point out the positive correlation between swimming speed, which partially depends on the torso and lower body muscle strength, and shot performance (Platanou and Varamenti). Finally, Ferragut Fiol et al. point out that there is a significant difference in the correlation between body measurements and shot performance among players in different positions (Ferragut Fiol et al.). This result means that future research that fails to account for player position might find it difficult to isolate the contributing factors and draw important insights about shot performance.

The following observations can be made about the contribution of different muscle groups to shot performance based on published literature. Prior research has not definitively identified the contribution of different muscle groups to ball speed and accuracy. Specifically, torso and body muscles have been studied extensively and results seem to be contradictory in the papers that they did examine them (McCluskey et al.; Platanou and Varamenti; Clarys et al.). There is no quantifiable evidence for the relative contribution of different muscle groups to shot accuracy and speed. From the reviewed papers, there is a confirmed correlation between shot performance and the lower body, torso, and upper body muscle activity intensity and muscle size in at least two studies (McCluskey et al.; Platanou and Varamenti), but at least one study disagrees for lower and torso muscle groups (Clarys et al.). According to water polo practitioners, stronger legs should equate to stronger shooting (Jayanti et al.; Solum) however given a lack of studies addressing this hypothesis and assertion, a lack of correlation tells us there is still a lot of research that needs to be conducted. Future research should aim to provide a definitive answer to this question.

The accepted but unproven consensus in answering what biomechanical factors influence water polo shooting, is that it is a combination of your entire body in sequence, however given a lack of scientific evidence to prove this accepted hypothesis true or not, it is difficult to make any definitive claims (Solum; Jayanti et al.). Published research has confirmed that trained players activate upper body muscle groups in sequence from distal to proximal (Yaghoubi, Esfehani, et al.) No results are reported for the position of the lower body muscle groups in the activation sequence. This result is consistent with the observation that experienced players release the ball at a greater angle and head height and with greater speed compared to inexperienced players (Melchiorri et al.).

Another important insight is that the position and experience plays a huge role in the biomechanics of water polo shots (Ferragut Fiol et al.; Yaghoubi, Esfehani, et al.). It is important to separate the study population according to player positions because even at the elite level, there are significant differences in the correlation between body measurements and shot performance among players in different positions.

A major limitation of this review is that there are significant differences in age, sex, physical conditioning and experience between the experimental subjects in the reviewed papers. For example, older athletes may be more experienced and therefore shoot harder or they may be out of shape and therefore slower than younger athletes. Male athletes at an equal competitive level to female athletes tend to shoot harder and may also shoot differently activating different muscles. Accordingly, research conclusions should not be directly compared across papers without further validation. As a result, none of my findings, which were indeterminate either way, can be used to draw definitive conclusions across all groups given the hyperspecificity of the experimental subjects in the studies conducted to date. In addition, future research should take into account position and experience in the experimental setups and methodology, as well as focus on largely under researched lower extremities in determining shooting power and the biomechanical factors influencing water polo shooting.

As a general observation, EMG studies seem to offer the deepest insights into muscle behavior, but it can be challenging to collect EMG measurements underwater (Papandrea). Exploring techniques to collect underwater EMG measurements from water polo shots is a promising future research topic. Another promising research approach is to examine EMG measurements and motion analysis at the same time to understand how they correlate so that we map activation patterns to muscle performance and, eventually, shot performance.

In conclusion, this review has shown that the existing research has barely scratched the surface of answering the question on which biomechanical factors influence a water polo skip shot. Consequently, further research in this area has the potential to uncover deeper biomechanical insights, in turn leading to attaining valuable competitive advantages.

Acknowledgements

Many thanks to Dr. Amanda Rounds for her mentorship during this investigation.

Works Cited

Clarys, J. P., et al. "An Electromyographic AMD Impact Force Study of the Overhand Water Polo Throw." *Biomechanics and Medicine in Swimming V1*, Taylor & Francis, 1992.

Feltner, Michael E., and Grant Taylor. *Three-Dimensional Kinetics of the Shoulder, Elbow, and Wrist during a Penalty Throw in Water Polo*. Aug. 1997. *journals.humankinetics.com*, https://doi.org/10.1123/jab.13.3.347.

Ferragut Fiol, Carmen, et al. "Water Polo Throwing Speed and Body Composition: An Analysis by Playing Positions and Opposition Level." *Journal of Human Sport and Exercise*, vol. 10, no. 1, 2015. *DOI.org (Crossref)*, https://doi.org/10.14198/jhse.2015.101.07.

Historical Anatomies on the Web: Bouglé, Julien Home. U.S. National Library of Medicine, https://www.nlm.nih.gov/exhibition/historicalanatomies/bougle_home.html. Accessed 26 Dec. 2024.

Hsu, Chiung-Yun, et al. "BIOMECHANICS ANALYSIS OF WATER POLO THROWING." *ISBS - Conference Proceedings Archive*, 2005. *ojs.ub.uni-konstanz.de*, https://ojs.ub.uni-konstanz.de/cpa/article/view/1145.

Jayanti, Vivin, et al. "AN ANALYSIS RUBRIC OF WATER POLO SHOOTING TECHNIQUE THROUGH BIOMECHANICS APPROACH." *Annals of Tropical Medicine & Public Health*, vol. 24, Jan. 2021. *ResearchGate*, https://doi.org/10.36295/ASRO.2021.24304.

McCluskey, Lisa, et al. "Throwing Velocity and Jump Height in Female Water Polo Players: Performance Predictors." *Journal of Science and Medicine in Sport / Sports Medicine Australia*, vol. 13, June 2009, pp. 236–40. *ResearchGate*, https://doi.org/10.1016/j.jsams.2009.02.008.

Melchiorri, Giovanni, et al. "Water Polo Throwing Velocity and Kinematics: Differences between Competitive Levels in Male Players." *The Journal of Sports Medicine and Physical Fitness*, vol. 55, Nov. 2014.

Papandrea, Patrizia. "The Use of Electromyography (EMG) Underwater." *Cometa Systems*, 6 Nov. 2024, https://www.cometasystems.com/the-use-of-electromyography-emg-underwater/.

Platanou, Theodoros, and Evdokia Varamenti. "Relationships between Anthropometric and Physiological Characteristics with Throwing Velocity and on Water Jump of Female Water Polo Players." *The Journal of Sports Medicine and Physical Fitness*, vol. 51, June 2011, pp. 185–93.

Solum, Jim. "FUNDAMENTALS OF THE WATER POLO SHOT." *Water Polo Planet*, https://www.waterpoloplanet.com/fundamentals-of-the-water-polo-shot/. Accessed 15 Oct. 2024.

Van den Tillaar, Roland. "The Biomechanics of the Elbow in Overarm Throwing Sports." *INTERNATIONAL SPORTMED JOURNAL*, vol. 6, Jan. 2005, pp. 7–24.

Yaghoubi, Mostafa, Mohamad Mahdi Esfehani, et al. "Comparative Electromyography Analysis of the Upper Extremity between Inexperienced and Elite Water Polo Players during an Overhead Shot." *Journal of Applied Biomechanics*, vol. 31, no. 2, Apr. 2015, pp. 79–87. *PubMed*, https://doi.org/10.1123/jab.2014-0068.

Yaghoubi, Mostafa, Amir Moghadam, et al. "Electromyographic Analysis of the Upper Extremity in Water Polo Players during Water Polo Shots." *International Biomechanics*, vol. 1, no. 1, Jan. 2014, pp. 15–20. *Taylor and Francis+NEJM*, https://doi.org/10.1080/23335432.2014.976591.

Sentiment Analysis for Youth Mental Welfare: A Comparative Study of Machine Learning Models By Vincent Qin

Abstract

Mental health concerns among youth are becoming increasingly prevalent, with 20% of United States adolescents experiencing mental health problems. A potential indicator of mental health concerns includes when a person's texts express overwhelming sadness or hopelessness. The study presents a comparison of methods to determine the emotional polarity of text. The models are trained on the Stanford SST2 and IMDb datasets as they are based on movie reviews, which exhibit particularly apparent emotions. The data is then encoded using a Bag-of-Words (BoW) strategy by only encoding the 10,000 most common words. Five models were tested in this paper: a decision tree, a random forest, the Adaboost classifier created with scikit-learn, a feedforward neural network with two hidden layers created using the PyTorch module, and a fine-tuned version of the model DistilBERT. The results are cross-validated by dividing the data into ten shards, training the model on nine shards, testing it on one shard, and then repeating this procedure for every shard. Finally, the models' accuracies were compared. The DistilBERT model had the highest overall accuracy (94.89%), which made it the most suitable for large-scale classification tasks. Due to the DistilBERT model's high learning (613 m) and inference times (38 ms), the slightly weaker neural network and Adaboost models are recommended for smaller-sized tasks. Although they have lower accuracies (88.79% and 80.21%, respectively), their short learning time (~1-2 hours) and inference times (<10 ms) are suitable for smaller tasks that can be manually verified.

Introduction

In 2021, the American Academy of Pediatrics, American Academy of Child and Adolescent Psychiatry, and Children's Hospital Association declared a national emergency in child and adolescent mental health. A youth's presence is heavily digital, consisting of texts, social media posts, and other interactions that can be logged. The National Institute of Mental Health finds that sadness and hopelessness are the most significant signs of depression. Consequently, these signs appear in their online activities. This paper seeks the best model to draw awareness to these signs of depression.

To conduct a fair comparison, several decisions must be made. The learning method will have the most significant impact on the results. This paper uses supervised learning as there are several free online datasets for sentiment analysis, such as the Stanford IMDb and SST2 datasets. This allows for closer observation of the model's performance as the data labels are known. The datasets above were chosen for three reasons. First, the datasets are based on movie reviews which display strong emotions and are classified into 0 (negative) or 1 (positive). Second, there are around 100,000 samples, which is more than enough to train a robust model. Third, everyone's online presence is different. Some might text in short messages with slang, while others might text very formally. It is the programmer's responsibility to consider this behavior,

and this paper did so by using two datasets. It is also beneficial that the class distribution of the datasets mimics that of real life. Next, it is crucial to decide what type of encoding to use. This paper discusses three main forms of encoding, namely: one-hot encoding, bag of words encoding, and transverse frequency-inverse document frequency encoding (TF-IDF). Each encoding type adds information to the resulting data vector and can highly impact the model's performance. Finally, the model choices should be decided upon. In this paper, decision trees, random forests, an implementation of the AdaBoost algorithm, feedforward neural networks, and a fine-tuned version of DistilBERT are all explored. Every model has its strengths and weaknesses. This paper's goal is to determine which one is best at classifying emotional polarity in large sets of text messages.

Code

The code for training the models and the datasets are linked on GitHub.

Datasets

Due to the complexity of the models, supervised learning methods were chosen. Although sentiment analysis models are traditionally trained on unlabeled data, supervised learning offers many advantages, the most prevalent of which is extracting patterns in the model's behavior to debug and improve the model's accuracy.

Everyone's messages are unique. Some people might be very formal and use proper punctuation, grammar, and sentence length. Others might utilize slang and divide their messages into many short texts. Therefore, the Stanford IMDb and Stanford SST-2 datasets were concatenated for a more balanced sample length spread. Since the data now had short and long sentences, a threshold of 100 words was appended as a new feature. These two datasets were chosen in particular because the emotional polarity distribution of online messages is very even. A study sampling 11 billion messages from the internet found that 6.51% were positive, 87.09% were neutral, and 6.40% were negative. Since the ratio of positive and negative classes is around 53:47, the data closely mirrored the emotional distribution of polarized digital communications.

Communication on the Internet is not always formal and long-winded. The models' accuracy was also tested on both human-made and AI-generated data. 30 short samples were labeled like the SST2 and IMDb Datasets. ChatGPT was used to create 170 samples for the AI-generated dataset.

Tokenization

It is important to divide the data into smaller characters, words, or sentences called tokens so the model can recognize patterns. This process is known as tokenization. Character-based tokenization was the first step. Every input was divided into its component characters using Python's .split() method. Letters, numbers, and punctuation were all divided into their own token. However, the model could not learn much from the singular characters. The model achieved an accuracy similar to random guessing due to the lack of information. Therefore, this

paper focuses more on word tokenization, which is superior to character-based tokenization since each word contains a meaning. The Natural Language Toolkit (NLTK) library's word_tokenize method was used to make each word, special character, number, and punctuation its token.

Encoding

Computers do not understand English. Instead, they must convert the words into numbers and perform operations on the numbers. To create these numbers, a crucial step known as encoding must be undertaken. This paper starts with one-hot encoding (OHE). In OHE, each word is represented by a binary vector. Every element is set to zero in the vector except for the index corresponding to the word. While this method is easy to implement, it has many limitations. In particular, the resulting matrices are massive, and the information is sparse when the vocabulary size is large. The model could not learn much from the data, and its accuracy was limited.

The Bag of Words (BoW) approach was used next, following one-hot encoding. Unlike one-hot encoding, BoW values the frequency of every word in the sample. The order of words is not saved. Each sample is represented by a vector where each element corresponds to the frequency of a word in the sample. BoW can better understand the importance of every word in the sample but can also result in large and sparse vectors if the vocabulary is large. It also fails to consider the relative importance of words across every sample. Even so, every model performed significantly better using BoW than OHE because BoW encodes orders of magnitude more information than OHE.

In response to the shortcomings of BoW, the paper also explored Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF improves BoW by considering the frequency of a word in a sample and how common that word is across all samples. This results in a value that emphasizes important words unique to a single sample while devaluing the common words across many documents. However, TF-IDF only achieved slightly better results than BoW. This is due to the complexity of the data. The large amount of words and samples makes the increase in information less significant, as the resultant vectors are still large and sparse.

A. Learning Algorithms

Since the paper attempts to classify the emotional polarity of a given sample, it is best to look into binary classifiers. Binary classification models work best for this task because their output is 0 or 1. They are also easy to train using the scikit-learn and PyTorch libraries and are computationally efficient.

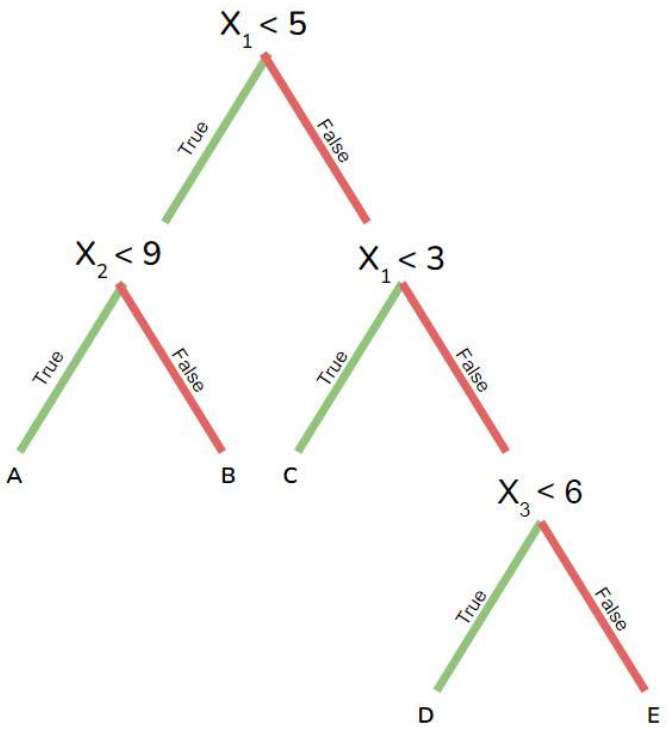


Figure 1: A flowchart simulating the processes of a decision tree

Decision tree classifiers operate by recursively splitting the dataset into subsets while trying to maximize a criterion. The two most commonly used criteria are:

1) Information Gain

Information Gain measures the decrease in entropy of a parent node and its children after a split. Splitting a dataset based on a feature shows how much information is gained. For a node x, its entropy is given by the formula

$$Entropy(x) = -\sum_{i=1}^{k} p_i log_2(p_i)$$

where p_i is the proportion of the sample in class i at node x and k is the number of classes. The information gain for a split is then

$$Information\ Gain = Entropy(parent) - \sum_{i=1}^{m} \frac{n_j}{n} Entropy(child_j)$$

where n_j is the number of samples in the child node j, n is the number of samples in the parent node, and m is the number of child nodes. Higher Information Gain indicates a better split. The split with the lowest entropy in child nodes is chosen.

2) Gini Index

The Gini Index is the measure of the impurity of a node. It quantifies how often a randomly chosen sample would be incorrectly labeled if it were randomly labeled based on the class distribution of the dataset. For a node x, the Gini Index is defined as

$$Gini(x) = 1 - \sum_{i=1}^{k} p_i^2$$

where p_i is the proportion of samples of class i at node x. The weighted average of the Gini Index of each split is then compared, and the split with the lowest weighted average is used.

The goal of a decision tree is to divide the data into subsets that contain the same label, which makes it comparatively simple to implement. By increasing the maximum depth, the number of splits of a decision tree can be increased. This subsequently increases its complexity. Due to the complexity of the data, however, the model cannot recognize patterns before it becomes prone to overfitting.

To counter overfitting, programmers often use pruning algorithms that remove the hyper-specific splits. This is often paired with a random forest classifier, a model that uses the aggregated vote of many decision trees. This paper does not implement a pruning algorithm because it risks underfitting the model by oversimplifying the model. Also, its impact would be minimal since its base model, a decision tree, is unable to make proper predictions at any depth. To have differently trained decision trees, the data is randomly sampled in a process known as bootstrap sampling to create multiple unique training subsets. Each subset is then fed into a decision tree. The final prediction is made by finding the majority vote across every tree. Random forests are generally more accurate and robust than singular decision trees, especially when working with large and noisy datasets.

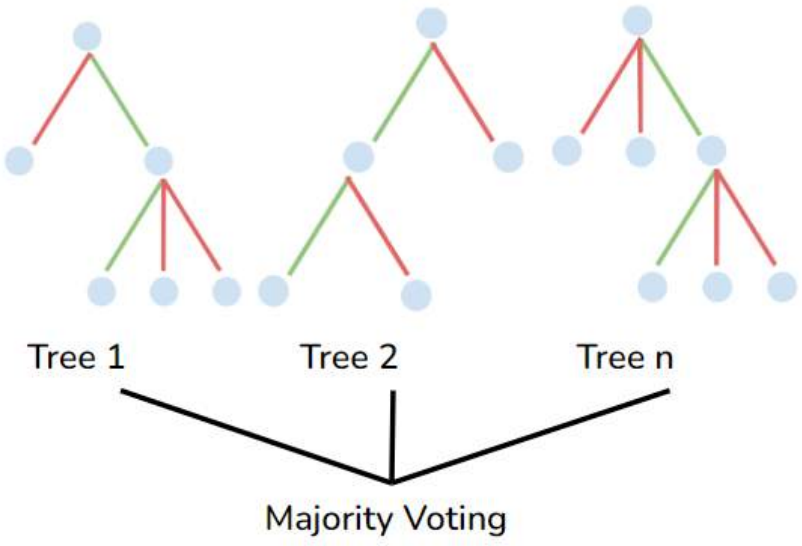


Figure 2: The voting process of a random forest

This paper implements the AdaBoost algorithm through the sklearn.ensemble library to

further improve the accuracy of the models. An AdaBoost classifier iteratively adds weak learners to a model to predict the difference between the predicted and target values. The model continues to fit weak learners until a stopping condition is met. Stopping criteria include a fixed number of iterations, a lack of improvement, and convergence of the loss function.

Another strong binary classifier is a neural network. This paper explores a Feedforward Neural Network specialized for binary classification tasks. The network consists of an input layer, two hidden layers, and an output layer. The input layer receives a vector representing the sample. Each hidden layer contains neurons that apply a linear transformation to the inputs followed by the Rectified Linear Unit (ReLU) activation function. The output layer has a single neuron with a sigmoid activation function, producing the probability that the sample is classified as 1. The network is trained using the binary cross-entropy loss function, which measures the difference between predicted and actual labels. This paper uses the AdamW optimizer, which updates the network edge weights based on the loss function to enhance training efficiency. A learning rate scheduler is also implemented to reduce the learning rate to fine-tune the training process periodically, helping to improve model performance.

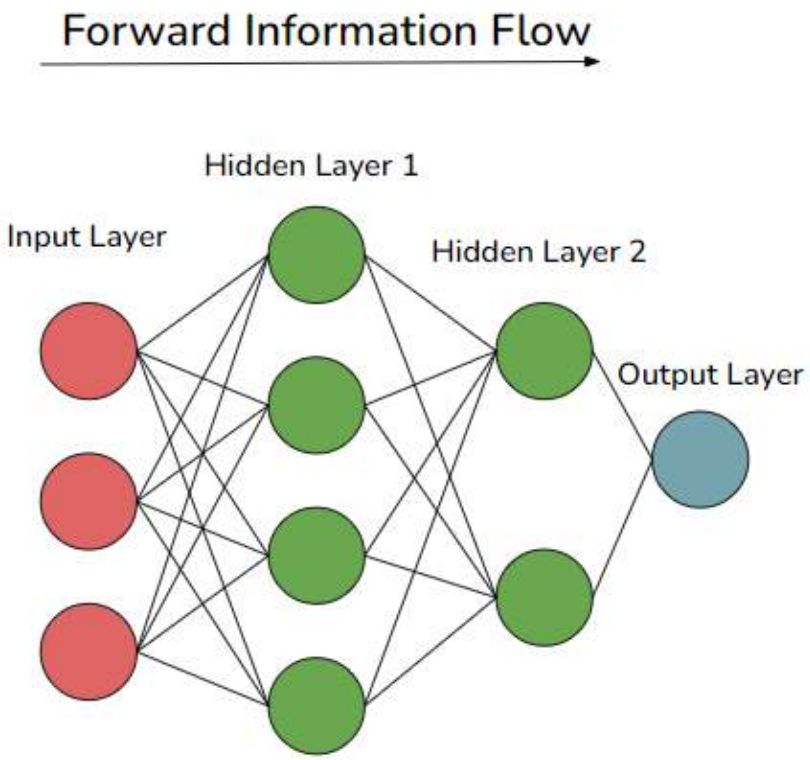


Figure 3: A model of a Feedforward Neural Network

This neural network structure is commonly used for binary classification tasks due to its simplicity and effectiveness in capturing complex patterns in the data.

DistilBERT is a smaller, faster, and lighter version of the BERT (Bidirectional Encoder

Representations from Transformers) model. It retains most of BERT's natural language processing abilities while being more efficient. BERT is trained on Wikipedia and the BookCorpus, and it has been trained on around 3.3 billion words in total. This allows BERT to develop a highly comprehensive understanding of language.

The fine-tuning process involves adjusting the weights of DistilBERT's edges based on the dataset. The binary cross-entropy loss function, which measures the difference between predicted and actual labels, must be optimized. The trainer class from the hugging face-transformers library manages data loading, optimization, and evaluation. The compute_metrics function calculates the accuracy of the model's predictions. However, DistilBERT would take nearly a week to be trained on the entire 100,000 sample dataset. To ensure a rigorous and timely training procedure, 10,000 samples were selected at random and were used to train DistilBERT. This setup allows the model to leverage its pre-trained language knowledge and adapt to the specifics of sentiment analysis, resulting in a robust and efficient classifier.

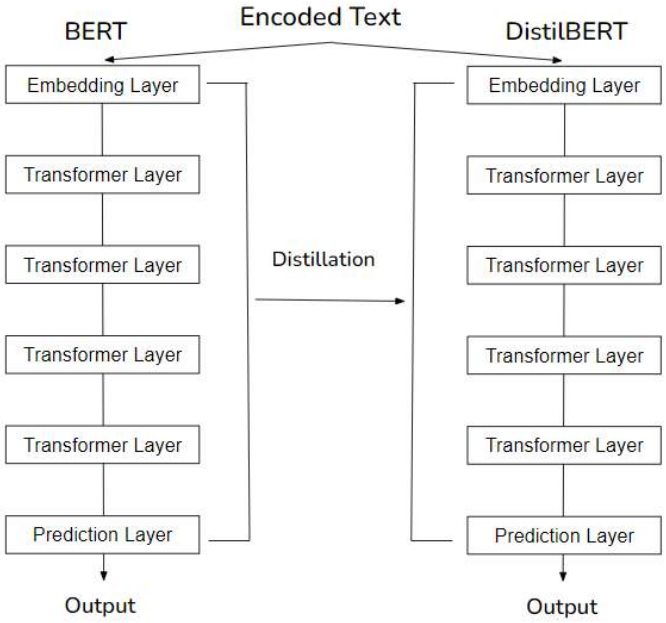


Figure 4: The creation process of BERT.

Validation

Validation techniques such as confusion matrices and cross-validation are employed to ensure the accuracy of the model. These methods help assess the model's performance and ensure it generalizes well to unseen data.

A confusion matrix is a powerful tool for evaluating the performance of a classification model. It provides a detailed breakdown of the model's predictions compared to the actual labels. The confusion matrix is as follows:

True Positive (TP): Model outputs 1, Actual label 1	False Positive (FP): Model outputs 1, Actual Label 0
False Negative (FN): Model outputs 0 Actual label 1	True Negative (TN): Model outputs 0, Actual Label 0

Figure 5: An example confusion matrix.

Using these, it is possible to calculate several important metrics, such as accuracy (1), precision (2), recall (3), and F1 Score (4).

$$\frac{(TP + TN)}{(TP + FP + FN + TN)} \quad (1)$$

$$\frac{(TP)}{(TP + FP)} \quad (2)$$

$$\frac{(TP)}{(TP + FN)} \quad (3)$$

$$\frac{TP}{TP+0.5(FP+FN)} \quad (4)$$

These metrics offer a comprehensive view of the model's performance and can be especially helpful when the data's class distribution is imbalanced.

Another validation method has to do with splitting the dataset. Instead of training and testing on one large dataset, the process first divides the data into ten subsets called folds. Then, using nine folds as the training set and the last fold as the validation set, the models are tested on their accuracy. The last step is to repeat this process ten times, using a different validation fold each time. This process is known as k-fold cross-validation. This helps ensure the model is not getting lucky in one fold. By averaging the accuracy across all the folds, the final accuracy can be determined. This ensures that the accuracy is representative of the model's true performance.

To determine if the models ran in a reasonable time, they were tested on a dataset of 1000 samples. Then, the Python time module was used to average the individual predictions. This helps ensure that the model does not take an absurdly long time to predict on a large dataset.

Using these techniques, the models' accuracy and reliability can be guaranteed.

Results: Character-Based Tokenization

Model	Accuracy	Loss
Decision Tree Classifier	52.41%	—

Table 1: Test on Character Based-Tokenization

The model performed poorly due to the information deficit in character-based tokenization. It could not decipher the patterns of singular characters, resulting in an accuracy

akin to blindly guessing.

Results: One-Hot Encoding

Model	Accuracy	Validation Loss	Learning Time (mins)	F1 Score	Inference Time (ms)
Decision Tree	54.87%	–	3	0.582	1
Random Forest	68.31%	–	43	0.741	4
Adaboost	70.73%	–	52	0.769	3

Table 2: Tests on One-Hot Encoding

One-hot encoding suffers a similar problem as character-based tokenization, where there is simply a need for more information. In this case, the encoded vectors are large and sparse. This makes it difficult for the models to find patterns in the data as the vectors have too little variety.

Results: Bag of Words Encoding

Model	Accuracy	Validation Loss	Learning Time (mins)	F1 Score	Inference Time (ms)
Decision Tree	62.87%	–	3	0.632	1
Random Forest	74.31%	–	43	0.741	6
Adaboost	79.44%	–	43	0.798	6
FNN	88.34%	0.638	126	–	9
DistilBERT	94.27%	–	526	–	36

Table 3: Tests on Bag of Words Encoding

Bag of Words encoding, on the other hand, adds more information into the encoded vectors and enables the models to achieve some strong predictions. However, it still suffers from many of the same issues as one-hot encoding.

Results: TF-IDF

Model	Accuracy	Validation Loss	Learning Time (mins)	F1 Score	Inference Time (ms)
Decision Tree	63.31%	–	3	0.643	1
Random Forest	74.25%	–	36	0.745	6
Adaboost	80.04%	–	49	0.809	5
FNN	89.29%	0.712	176	–	9
DistilBERT	94.60%	–	533	–	38

Table 4: Tests with TF-IDF.

Although TF-IDF mitigates most of BoW's flaws, there is only a slight increase in accuracy. This could indicate that the accuracy is limited by the model design or dependent on factors outside of encoding.

Results: Other Datasets

Model	Accuracy - AI Data	Accuracy - Human Data
Adaboost	79.56%	80.12%
FNN	89.34%	88.79%
DistilBERT	94.78%	94.45

Table 5: Tests with human and AI-generated data.

There was no significant difference in the model's ability to predict formal datasets, my dataset, or the ChatGPT dataset.

Discussion

A fine-tuned version of DistilBERT is the strongest model. It achieves a 94.89% accuracy which is extremely accurate and can be used in almost every task. Since the models are to be implemented to scan large libraries of texts and messages, it would be best to use the most accurate model. The main drawback of DistilBERT is its high inference and learning times. Learning time is the time it takes to train the model. It is not a big issue as the model only needs to be trained once. It is only inconvenient when the model is updated because re-training is only required under such circumstances. To minimize this issue, it is recommended that the number of epochs or the size of the dataset be decreased. Inference time, on the other hand, matters a lot when dealing with large amounts of data. To predict on a dataset of a billion samples, the

DistilBERT model would require a long time. The high accuracy, however, makes such a trade-off worth it, as it eliminates the need for manual verification.

The decision tree classifier stands out as the most intuitive model. However, it is difficult to make many meaningful splits without overfitting. For smaller tasks, either the neural network or the Adaboost algorithm is recommended. Both have sufficient accuracy, and a small use case would ensure that it is possible to verify each prediction. To improve Adaboost, using the algorithm on the predictions of a random forest works significantly better than that of a sole decision tree. This is due to the complexity of the data, as a single decision tree is too weak for a learner. One can change the size and number of layers to improve the neural network. Currently, it is prone to overfitting on the training dataset, and reducing the complexity of the neural network slightly could improve its validation loss score.

Conclusion

Among the many models tested in this paper, DistilBERT stands out as the most accurate. To be used on a massive collection of text, increasing its accuracy will reduce the need for human verification. By leveraging DistilBERT's advanced capabilities for sentiment analysis, it is possible to more effectively detect signs of mental health issues from digital communication and raise awareness. The California Department of Healthcare Services finds raising awareness can reduce the stigma associated with receiving treatment for mental illnesses. This paper aims to aid many of the 280 million people affected by depression by bringing attention and directing support to those in need.

Although movie reviews are a good indicator of emotion, meaning might be sometimes concealed. In particular, a text might contain sarcasm, which is difficult for machine learning models to handle. Furthermore, the degree of emotion also varies drastically with every message, so a slightly negative text might be processed differently from an extremely negative text. The challenge is ensuring the model can discern the text's true meaning. To do so, the model has to be trained again to understand sarcasm, slang, and the other complexities of language.

Acknowledgments

I would like to thank my mentor, Jason Liang, for guiding me through the research process and my family for supporting me.

Works Cited

American Academy of Pediatrics. "AAP-AACAP-CHA Declaration of a National Emergency in Child and Adolescent Mental Health." *Www.aap.org*, American Academy of Pediatrics, 19 Oct. 2021, www.aap.org/en/advocacy/child-and-adolescent-healthy-mental-development/aap-aacap-cha-declaration-of-a-national-emergency-in-child-and-adolescent-mental-health/. Accessed 18 July 2024.

Ansel, Jason, et al. *PyTorch 2: Faster Machine Learning through Dynamic Python Bytecode Transformation and Graph Compilation*. 27 Apr. 2024, https://doi.org/10.1145/3620665.3640366.

Berrar, Daniel, and Werner Dubitzky. "Information Gain." *Springer EBooks*, 1 Jan. 2013, pp. 1022–1023, https://doi.org/10.1007/978-1-4419-9863-7_719.

Bird, Steven, et al. *Natural Language Processing with Python*. Beijing Etc., O'reilly, 2009.

Dodge, Yadolah. "Gini Index." *The Concise Encyclopedia of Statistics*, 2021, pp. 231–233, link.springer.com/referenceworkentry/10.1007%2F978-0-387-32833-1_169, https://doi.org/10.1007/978-0-387-32833-1_169.

GeeksforGeeks. "Bellman–Ford Algorithm | DP-23 - GeeksforGeeks." *GeeksforGeeks*, Dec. 2012, www.geeksforgeeks.org/bellman-ford-algorithm-dp-23/.

Maas, Andrew L., et al. "Learning Word Vectors for Sentiment Analysis." *ACLWeb*, Association for Computational Linguistics, 1 June 2011, www.aclweb.org/anthology/P11-1015. Accessed 17 June 2024.

"Mental Health Awareness." *Ca.gov*, 2023, www.dhcs.ca.gov/services/MH/Pages/MHAM_Matters.aspx.

National Institute Of Mental Health. "Depression." *National Institute of Mental Health*, Mar. 2023, www.nimh.nih.gov/health/topics/depression. Accessed 23 July 2024.

National Library of Medicine. *Child and Adolescent Mental Health*. *Www.ncbi.nlm.nih.gov*, Agency for Healthcare Research and Quality (US), 1 Oct. 2022, www.ncbi.nlm.nih.gov/books/NBK587174/.

Pedregosa, Fabian, et al. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research*, vol. 12, no. 85, 2011, pp. 2825–2830, jmlr.csail.mit.edu/papers/v12/pedregosa11a.html.

Refaeilzadeh, Payam, et al. "Cross-Validation." *Encyclopedia of Database Systems*, 2009, pp. 532–538, link.springer.com/referenceworkentry/10.1007%2F978-0-387-39940-9_565, https://doi.org/10.1007/978-0-387-39940-9_565.

Sanh, Victor, et al. "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter." *ArXiv.org*, 2019, arxiv.org/abs/1910.01108.

Shultz, Thomas R., et al. "Confusion Matrix." *Encyclopedia of Machine Learning*, 2011, pp. 209–209, https://doi.org/10.1007/978-0-387-30164-8_157.

Socher, Richard, et al. "Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank." *ACL Anthology*, Oct. 2013, pp. 1631–1642, www.aclweb.org/anthology/D13-1170. Accessed 18 June 2024.

Uther, William, et al. "TF–IDF." *Encyclopedia of Machine Learning*, 2011, pp. 986–987, https://doi.org/10.1007/978-0-387-30164-8_832.

How Collectivism Bolsters Japan's Metabolic Health: Lessons For The U.S
By Caitlyn Zhu

Abstract

The United States has a poor life expectancy and high spending on healthcare, in large part due to an epidemic of metabolic disease. Despite a weaker economy and less spending on healthcare, Japan has a very low rate of metabolic disease and greater life expectancy. Explaining this contrast, this essay explores several ways in which Japan's vertical collectivism improves its metabolic health. In Japan, vertical collectivism is responsible for centralized education, higher taxes, and high quality public transportation. The resulting economic and activity equality makes a healthy diet and regular exercise accessible to many people, thereby reducing metabolic disease in Japan. Comparing the United States and Japan from this angle reveals that social and economic inequality severely impact public health, while well-designed public services may reduce the vulnerability of those in poverty to metabolic disease. The lessons in public health that Japan provides could be valuable in increasing lifespan and reducing healthcare spending in the United States.

Introduction

Despite being an economic powerhouse, the United States has a notoriously low life expectancy of 78.8 years as of 2019. (CDC, National Center for Health Statistics) At the same time, the country spends several trillions of dollars a year on healthcare, or 17.3% of the country's GDP in 2022. (American Medical Association, 2024) Reforms in American health must be sought to better American health while cutting down on spending. In contrast, Japan is known for having one of the longest life expectancies of any country, at 84.3 years as of 2019. (WHO, "Health data overview for Japan"). It achieves this while spending only 11% of its GDP on healthcare as of 2021. Metabolic and dietary health is one major reason for this vast disparity. This essay aims to explore social and cultural forces behind the gap.

Metabolic disease is one of the United States' most pressing public health problems and a significant source of preventable, expensive disease. Poor nutrition accounts for 45% of annual deaths from stroke, heart disease, and diabetes. (Micha et. al, 2017) The 2017-2018 National Health and Nutrition Examination Survey found that about 42.5% of United States adults ages 20 and over are obese. Obesity accounts for hundreds of billions of dollars spent on American healthcare per year. (Cawley et. al, 2021) However, health problems from poor nutrition also extend far beyond the obese population. Araújo et. al compared blood glucose, blood pressure, HbA1c, HDL-C, and triglycerides with acceptable levels reported by the 2001 National Cholesterol Education Program Adult Treatment Panel III, finding that only 19.9% of Americans were metabolically healthy. (2021) This means that many normal weight Americans also suffer from metabolic dysfunction, encompassing conditions such as elevated blood pressure and insulin resistance, which increase the risk of deadly chronic diseases.

The rate of obesity among Japanese adults ages 20 and over is a mere 3.8% for males and 3.2% for females. (Nobuo Yoshiike, 2013) Many papers have explored how the traditional Japanese diet contributes to this low obesity and high lifespan of 84.5 years. (World Health Organization) This diet is based on white rice, seafood, pickled vegetables, and soy products such as miso and tofu. It contains less fat, sugar, red meat, dairy, and grains than the American diet. Although most Japanese no longer conform strictly to a traditional diet, data from the Food and Agriculture Organization of the United Nations shows that Japan still consumes less sugar and more seafood and soybeans than other countries.

Crucially, ultra-processed foods, which lack vital nutrients and contain added sugar and preservatives, have infiltrated the Japanese diet to a lesser extent than the United States. A 2019 study found that ultra-processed foods account for 38.2% of the caloric intake of Japanese aged 30-59. (Kaori Koiwai et. al, 2019) It is estimated that over a quarter of the caloric intake of Japanese youth comes from these foods. (Shinozaki et. al, 2024) Yet an astounding 57.9% of caloric intake for people ages 2 and above in the United States comes from ultra-processed foods. (Eurídice Martínez Steele, 2015) Although not all ultraprocessed foods are injurious, the vast majority of them cause disease from added sugar. By taking up space from more nutritious foods in the diet, they also contribute to malnutrition. Their consumption is associated not only with obesity and diabetes, but heart disease and stroke, spelling trouble for public health. (Leonie Elizabeth et. al, 2020) (Jennifer M. Poti et. al, 2018)

This essay examines one fundamental difference between the two countries that has significant implications for metabolic health: Japan's collectivist culture. Collectivism is broadly defined as the idea that the needs and desires of the individual are subordinate to those of the collective it belongs to. A study found that among a sample of 53 countries, individualism was associated with higher BMI. (Masood et. al, 2019) A 2022 study using data from 155 countries confirmed this result for men. (Akaliyski et. al) However, this topic has not yet been explored in detail. Collectivism shapes Japan's economic, educational, medical and political systems. It strengthens societal norms of harmony and conformity among individuals, groups, businesses, and institutions. Japan's collectivism is also highly hierarchical, leading to the centralization of power and greater compliance with power. This enables the national government to intervene in business, education, and health more successfully than in the United States. This essay examines economic equality, public education, public transportation, and lastly healthcare, each a mechanism by which Japanese metabolic health is improved by hierarchical collectivism. By comparing the United States and Japan through this lens, we gain insight into how a government's public health policies interact with culture and how to make these policies more effective.

Collectivism, economic inequality, and nutrition

In the United States, many studies have found that people in poverty are at greater risk of obesity. (Wada et. al, 2016) People in poverty are inclined to consume more ultra processed foods because of their caloric density and affordability, and also have less access to recreation.

This phenomenon, called the obesity poverty paradox, is also exhibited by European countries. (Wioletta Żukiewicz-Sobczak et. al, 2014) Japan appears to defy the paradox. The country has yet to recover from the "Lost Decade" that followed Japan's stock market crash in 1990. With poverty defined by income below half of the median, Japan has the highest poverty rate among OECD countries at 15.4% as of 2018. ("Poverty rate") In theory, it should have high obesity to match; and yet obesity remains far less common in Japan than in the United States.

Rather than absolute poverty, degree of economic inequality may prove a more potent cause of obesity. In the United States, economic inequality is high due to a number of factors including low public welfare and the perpetuation of historical racial inequality. (Timothy M. Smeeding, 2005) This may be a reason that obesity is so high, because inequality reduces the number of people able to afford a healthy lifestyle. One study in New York State found that economic inequality on the county level, as measured by a Gini coefficient, is associated with higher prevalence of obesity in women. (Kim et. al, 2018) It may be because hierarchical collectivism reduces economic inequality that the effect of poverty is less deleterious in Japan than in the United States.

In Japan, the fruits of economic development during the Meiji Restoration were redistributed across the country by the Liberal Democratic Party (LDP). In governing policy consistent with collectivist culture, the LDP used money from large companies to subsidize farmers and small businesses. This "iron triangle," composed of the LDP, large companies, and farmers and small businesses, resulted in high economic equality in Japan because vulnerable social groups were not left behind during a critical period of industrialization. (Andrew Gordon, 279) Recently, income inequality has been on the rise since Japan's economy slumped in 1990. For example, while income remains relatively equal among full-time employees, these employees represent a shrinking portion of the workforce. Part-time workers receive only 40% the wage per hour that full-time workers do. (Marco Mira d'Ercole, 2006) Still, Japan enjoys relatively high economic equality compared with the United States, one of the most unequal among the world's developed nations. For example, the most recent Gini index measurements are 39.8 for the United States and 32.9 for Japan (The World Bank, "Gini index - OECD members").

Japanese social institutions bolster economic equality throughout childhood and adulthood. School curriculum and financing are highly centralized, resulting in even distribution of educational resources across the country. (Merry I. White, 99) Equality in education contributes to more equal job opportunities and less future income inequality for students. Next, when Japanese students become adults, they also face higher tax rates in Japan, further reducing economic inequality. The inheritance tax is 10-55%, preventing wealth from accumulating in small portions of the population over generations. (Koike Yuriko, "Why inequality is different in Japan") In the United States, funding and curriculum for schools is not nearly as standardized, mostly derived from state funding and property taxes as opposed to federal funding. This results in great variation in the quality of education across geographical locations. Next, inheritance taxes are extremely lax, allowing wealth to accumulate. These problems allow historical

inequalities, such as those from American slavery, to be perpetuated to the current day. Comparative economic equality in Japan supports metabolic health.

National school lunch and Japanese nutrition in childhood and beyond

Youth is a formative period in many ways including with respect to health. Metabolically healthy children are more likely to grow into metabolically healthy adults. On this front, American national school lunch has a vital role to play. National Health and Nutrition Surveys from 2007-2012 reveal that U.S students aged 5-18 receive 47% of their daily caloric intake from National School Breakfast and School Lunch Program meals. (Karen Weber Cullen, 2017) Low-income students especially are reliant on school meals for nutrition. Unfortunately, American school lunch usually contains many ultra processed foods. The federal government provides grants to schools for lunch, but the meals themselves are sourced from private companies. (Nicholas Confessore, "School Lunch") This results in greater variation in the quality of meals across schools and an overall reduced standard of wholesomeness. A study in 2005 found that students on reduced-price school lunch had more obesity because they consumed about 60 extra calories a day compared to those who ate lunch from home. (Schanzenbach, 2005) The prioritization of palatability and food companies' profit has also resulted in lax nutritional requirements for American school lunches. For instance, potatoes and tomato sauce can fulfill required servings of vegetables. To fight childhood obesity, the Obama administration passed the Healthy, Hunger-Free Kids Act of 2010, which strengthened regulations about the quantity of fruit, vegetables, sodium and fat in school lunches. A 2023 study found that the law resulted in significant decreases in BMI for low-income children. (Aruna Chandran) Among this group, it reduced childhood obesity prevalence by 47% by 2018. (Erica L. Kenney)

In Japan, national lunches are designed to promote the metabolic health of all children. Like in the United States, lunches for elementary and junior high students are paid for by the national government. Unlike in the United States, under its collectivistic culture, school lunches in Japan are of standardized quality, designed by nutritionists and often prepared using local ingredients. A 2017 study concluded that nutrient-wise, Japanese school lunches made an overall positive contribution to the diets of Japanese children. (Keiko Asakura)

However, the most remarkable way that Japanese national school lunch diverges from the United States is its educational aspects. Japanese schools are expected to take responsibility for the discipline and health of students. Since the 2005 Basic Law on *Shokuiku*, nutrition experts in Japanese schools were not only to work in the background designing meals, but directly teach children about nutrition (Teiji Nakamura, 2008). The goals of this law were to reduce metabolic disease, promote the traditional Japanese diet, and improve eating habits. Additionally, students are taught table manners, and encouraged to eat all the food they are served so that consumption does not rely solely on palatability as in America. This high-involvement national lunch program reflects the collectivist ideology characteristic of Japanese elementary and middle school classrooms. It homogenizes Japanese youth in nutritional literacy regardless of economic status, improving individuals' ability to lead a healthy lifestyle. High quality school lunches and

nutritional education are likely one reason that Japan consumes less ultra processed foods than America, improving the former's overall metabolic health.

Hierarchical collectivism is at the root of the sentiment that the school, or any social institution, is responsible for the health of individuals. In Japan, not only do schools play a role in the health of children, companies also play a role in the health of adults. In 2008 the Japanese government passed a "Metabo" law stating that companies whose workers had metabolic disorder would be fined if they did not reduce metabolic disorder to 10% among their employees by 2012. ("Japanese firms face penalties for overweight staff") Since then companies have been required to annually weigh and measure the waist circumference of their employees. In the United States this policy, as well as lunchtime discipline for students, would be deemed violations of personal freedom. Yet in Japan, hierarchical collectivism reinforces the sentiment that failure in individual health is also the failure of the company, the school, and the state. It is for this reason that these institutions can implement strict, uniform rules upon people's health through *Shokuiku* and "Metabo."

Public transportation and metabolic health

After nutrition, the second reason that the Japanese have healthier metabolisms is exercise, which lowers blood sugar by increasing insulin sensitivity. Like nutrition, the state again strongly influences Japanese exercise, which may be an unintended, but no less important, benefit of the country's high quality public transportation. Japan's reliable public transportation reflects centralization of power to the government: one group of state-managed railway companies called Japan Railways provides most of Japan's trains. (Geradi Yudhistira, 2015) Its widespread use also reflects collectivist culture, as personal transportation is more sought after in an individualistic society. By using this service instead of driving cars, the Japanese incorporate more walking into their schedule by going to and from public stations. (Naoki Kondo, 2014) As a result, Americans walk about 5,000 steps a day on average, while Japanese males and females ages 20 and above walk 6,793 steps and 5,832 steps a day respectively. (Emily Washburn, "Daily Step Targets For Americans") (Hirokage Tabata, "Walking Reduces Mortality Rate")

The wide availability of public transportation may be even more beneficial than the additional exercise in and of itself. A 2017 study, using smartphone data from 111 countries, found that countries' Gini coefficient for activity inequality was a better predictor of obesity levels than average activity. (Tim Althoff et. al, 2017) This is because, critically, low activity individuals benefit significantly more from additional activity than high activity individuals. In Japan, public transportation improves activity equality because it is a source of exercise accessible to the whole population. American public transit is notoriously poor quality, so transportation is not often a reliable source of exercise. Access to recreational spaces and equipment, which are both lower for people in poverty, are more important determinants of activity level. (Brownson et. al, 2003) Hence Japan's Gini coefficient for activity is 0.248, whereas America's is 0.303. (Tim Althoff et. al, 2017) Activity inequality contributes to obesity much in the same way of economic inequality.

Lastly, it is worth noting that aside from providing a source of exercise, good public transportation reinforces economic equality because it provides broad access to schools and jobs. By increasing access to grocery stores it also mitigates the risk of food deserts, or areas without ready access to nutritious and fresh food, which have long been observed plaguing the United States. Nevertheless, mainly due to the economic recession, food deserts do exist in Japanese cities as well. (Ikejima, 2015)

Food and healthcare companies

American pharmaceutical companies tout biotechnology as a solution to the nation's obesity epidemic. But low pharmaceutical innovation appears not to have hindered Japan's metabolic health. From 1992-2004, Japan produced 9.7% of the world's new molecular entities (NMEs) or drugs not previously approved by the FDA, while the United States produced 43.7% of them. (Salomeh Keyhani, 2010) Collectivist countries like Japan are known to innovate less, because a conformist environment is less conducive to new or unusual ideas. (Johannes C. Buggle, 2020) In the United States, individualist culture primes an innovative environment, which drives a multi-billion dollar pharmaceutical business. (Mark Zachary Taylor et. al, 2012) The reason Japan still has a years-longer lifespan is that on the scale of a nation, obesity and associated deadly diseases are social ailments caused by the distribution and quality of basic resources. Tirzepatide injections will not move the needle on American lifespan if children across the nation are still served modified Domino's pizzas for lunch. Prevention ultimately saves more human lives, suffering, and money than treatment. Policies that reduce social inequality, either directly through wealth taxes or indirectly like *Shokuiku*, often present more cost-effective solutions for metabolic health than does medicine.

The American healthcare system provides no safety net for metabolic disease perpetuated by American social and economic systems. In 2023, 65.4% of healthcare coverage was provided by private companies. (United States Census, 2024) As a result, much of the 4.9 trillion that the United States spent on healthcare was wasted on administration, reducing both affordability and timeliness of treatment. In 2019, 1 out of 4 Americans with type 1 or type 2 diabetes reported reducing insulin use due to the cost. (Tseng et. al, 2020) In comparison, healthcare is much more accessible in Japan through universal healthcare. According to Japan's copayment system, 30% of health costs are paid by users while 70% is covered by the government. Insulin and other prescription drugs are much more widely accessible due to caps on profit margins for drugs covered by mandatory insurance. In 2021, the price of insulin per vial was $14.40 in Japan and $98.70 in the United States. (Oladimeji Ewumi, "The High Costs of Insulin") Far from being a real solution to diabetes, medicine is not accessible to the Americans most vulnerable to metabolic diseases.

Possible drawbacks of collectivism in metabolic health

Collectivism confers major benefits to metabolic health overall, but may not be an unqualified positive. Due to conformist culture in Japan, groups who are different from the

majority, including those with BMI or other metabolic disorders, are stigmatized. This attitude is promoted by legislation like "Metabo." Therefore, the metabolic health of the Japanese may come with an associated cost to mental health. It is also possible that stress from social shaming can counterproductively hinder the improvement of physical and metabolic health.

Next, nutritional education in schools has thus far limited itself to a traditional Japanese diet, since Japan is very ethnically homogeneous with 98.4% of its population being Japanese (Yasuo Masai, "People of Japan"). It is predicted that by 2030, Japan will need 4.19 million foreign residents, nearly a million more than in 2023, to achieve its GDP growth targets. ("Japan will need 4 times as many foreign workers") Because of Japan's aging crisis, it will likely prove necessary for the national government to encourage more foreigners to settle in Japan for the long-term. If larger foreign populations become permanently integrated in Japan, nutritional education that is inclusive of various cultures may become important to the continued success of *Shokuiku*.

Conclusion

An individual's path to disease is heavily contingent on the social institutions shaping their living conditions. Metabolic disorder is a societal ailment which cannot be cured solely with medical treatments prescribed to individuals. Hierarchical collectivism is the basis for Japan's good nutritional health because it bolsters the government's power to centralize the country's resources and tightly regulate public services. This results in high quality public education and transportation, high taxes, and other conditions which reduce social, economic and even activity inequality, likely alleviating the impact of Japan's high poverty on metabolic health. In contrast to Japanese social institutions, American pharmacies and food companies fail to address the fact that public health belongs to the collective.

American policymakers are currently concerned with how public health can be improved within the context of American legal and social systems. The Japanese harness conformity in their public health programs. Could the United States harness diversity to the same end? What is the balance between conformity and the individual right to health? This essay suggests the salience of these questions even as it does not address them.

As for a single concrete takeaway for the United States from Japan, reform in national school lunch is an accessible public health intervention that would have a high impact. Increased involvement of nutritionists in designing American lunch menus may greatly benefit children's wellbeing. As the Healthy, Hunger-Free Kids Act of 2010 demonstrated, nutritious school lunches disproportionately help low-income students, who are at higher risk of malnutrition and metabolic disease and often rely heavily on school lunches. Nutritious lunches improve attendance and performance in school while also being cheaper than reducing class sizes, so they play a role in breaking the cycle of poverty. ("Quality of school lunch") Ultimately, any public health intervention that helps prevent metabolic disease and poverty would result in savings on healthcare for the United States.

Broadly, the Japan-United States comparison reveals one valuable guiding principle about public health. In a country with a developed economy, the ailment that most damages public health the most is inequality. Government intervention promotes the equal distribution of resources which support metabolic health. In particular, well designed public services in education and transportation have potential to mitigate the poor metabolic health associated with poverty. In the United States, further investment in such services would likely have great returns in the form of increased lifespan and productivity at work with lower healthcare costs. Given the global epidemic of metabolic disease that is worsening by the year, the valuable lesson that Japan provides from vertical collectivism and public health is especially salient.

Works Cited

National Center for Health Statistics (CDC), "United States. Life Expectancy Increased in 2019, Prior to the Pandemic," 2020 https://www.cdc.gov/nchs/pressroom/nchs_press_releases/2020/202012.htm Accessed June 2024

American Medical Association, "Trends in healthcare spending," updated July 2024 https://www.ama-assn.org/about/research/trends-health-care-spending Accessed January 2025

World Health Organization, "Health data overview for Japan" https://data.who.int/countries/392 Accessed June 2024

Micha, Renata et al. "Association Between Dietary Factors and Mortality From Heart Disease, Stroke, and Type 2 Diabetes in the United States." *JAMA* vol. 317,9 (2017): 912-924. doi:10.1001/jama.2017.0947

Cawley, John et al. "Direct medical costs of obesity in the United States and the most populous states." *Journal of managed care & specialty pharmacy* vol. 27,3 (2021): 354-366. doi:10.18553/jmcp.2021.20410

Araújo, Joana et al. "Prevalence of Optimal Metabolic Health in American Adults: National Health and Nutrition Examination Survey 2009-2016." *Metabolic syndrome and related disorders* vol. 17,1 (2019): 46-52. doi:10.1089/met.2018.0105

Yoshiike, Nobuo, and Miki Miyoshi. *Nihon rinsho. Japanese journal of clinical medicine* vol. 71,2 (2013): 207-16.

World Health Organization (WHO), "Japan" https://data.who.int/countries/392 Accessed January 2025

Koiwai, Kaori et al. "Consumption of ultra-processed foods decreases the quality of the overall diet of middle-aged Japanese adults." *Public health nutrition* vol. 22,16 (2019): 2999-3008. doi:10.1017/S1368980019001514

Shinozaki, Nana et al. "Highly Processed Food Consumption and its Association With Overall Diet Quality in a Nationwide Sample of 1,318 Japanese Children and Adolescents: A Cross-Sectional Analysis Based on 8-Day Weighed Dietary Records." *Journal of the Academy of Nutrition and Dietetics*, S2212-2672(24)00267-3. 7 Jun. 2024, doi:10.1016/j.jand.2024.06.001

Martínez Steele, Eurídice et al. "Ultra-processed foods and added sugars in the US diet: evidence from a nationally representative cross-sectional study." *BMJ open* vol. 6,3 e009892. 9 Mar. 2016, doi:10.1136/bmjopen-2015-009892

Elizabeth, Leonie et al. "Ultra-Processed Foods and Health Outcomes: A Narrative Review." *Nutrients* vol. 12,7 1955. 30 Jun. 2020, doi:10.3390/nu12071955

Poti, Jennifer M et al. "Ultra-processed Food Intake and Obesity: What Really Matters for Health-Processing or Nutrient Content?." *Current obesity reports* vol. 6,4 (2017): 420-431. doi:10.1007/s13679-017-0285-4

Masood, M., Aggarwal, A. & Reidpath, D.D. Effect of national culture on BMI: a multilevel analysis of 53 countries. *BMC Public Health* 19, 1212 (2019). https://doi.org/10.1186/s12889-019-7536-0

Akaliyski, Plamen et al. "The weight of culture: Societal individualism and flexibility explain large global variations in obesity." *Social science & medicine (1982)* vol. 307 (2022): 115167. doi:10.1016/j.socscimed.2022.115167

Żukiewicz-Sobczak, Wioletta et al. "Obesity and poverty paradox in developed countries." *Annals of agricultural and environmental medicine : AAEM* vol. 21,3 (2014): 590-4. doi:10.5604/12321966.1120608

Organisation for Economic Co-operation and Development (OECD), "Poverty rate" https://www.oecd.org/en/data/indicators/poverty-rate.html Accessed January 2025

Timothy M. Smeeding, 2005. "Public Policy, Economic Inequality, and Poverty: The United States in Comparative Perspective," Social Science Quarterly, Southwestern Social Science Association, vol. 86(s1), pages 955-983, December.

Kim, Daniel et al. "Geographic Association Between Income Inequality and Obesity Among Adults in New York State." *Preventing chronic disease* vol. 15 E123. 11 Oct. 2018, doi:10.5888/pcd15.180217

Andrew Gordon, A Modern History of Japan, Fourth Edition, Oxford University Press 2008, 279.

Marco Mira d'Ercole, "Income Inequality and Poverty in OECD Countries: How Does Japan Compare?" The Japanese Journal of Social Security Policy, Vol.5, No.1 (2006)

The World Bank, "Gini index - OECD members" https://data.worldbank.org/indicator/SI.POV.GINI?locations=OE&most_recent_value_desc=false Accessed June 2024

Merry I. White, The Japanese Educational Challenge: A Commitment to Children, Touchstone 1988, 99.

Koike Yuriko, "Why Inequality is Different In Japan," World Economic Forum, 2015 https://www.weforum.org/agenda/2015/03/why-inequality-is-different-in-japan/

Cullen, Karen Weber, and Tzu-An Chen. "The contribution of the USDA school breakfast and lunch program meals to student daily dietary intake." *Preventive medicine reports* vol. 5 82-85. 28 Nov. 2016, doi:10.1016/j.pmedr.2016.11.016

Nicholas Confossore, "How School Lunch Became the Latest Political Battleground," The New York Times, 2014, http://nyti.ms/1xZQvn8

Diane Whitmore Schazenbach, "Do School Lunches Contribute to Childhood Obesity?" University of Chicago, 2005 https://academics.hamilton.edu/economics/home/lunch6.pdf

Chandran, Aruna et al. "Changes in Body Mass Index Among School-Aged Youths Following Implementation of the Healthy, Hunger-Free Kids Act of 2010." *JAMA pediatrics* vol. 177,4 (2023): 401-409. doi:10.1001/jamapediatrics.2022.5828

Erica El Kenney et. al, "Impact of the Healthy, Hunger-Free Kids Act on Obesity Trends," Health Affairs, Vol. 39, No. 7, 2020, https://doi.org/10.1377/hlthaff.2020.00133

Asakura, Keiko, and Satoshi Sasaki. “School lunches in Japan: their contribution to healthier nutrient intake among elementary-school and junior high-school children.” *Public health nutrition* vol. 20,9 (2017): 1523-1533. doi:10.1017/S1368980017000374

Nakamura, Teiji. “The integration of school nutrition program into health promotion and prevention of lifestyle-related diseases in Japan.” *Asia Pacific journal of clinical nutrition* vol. 17 Suppl 1 (2008): 349-51.

Justin Mccurry, The Guardian, “Japanese firms face penalties for overweight staff,” 2009 https://www.theguardian.com/world/2008/mar/19/japan#:~:text=Corporate%20Japan%20will%20join%20the,their%20employees'%20weight%20under%20control Accessed January 2025

Geradi Yudhistira et. al, “Transportation System in Japan: A Literature Study,” Jurnal Manajemen Transportasi & Logistik, 2015, http://dx.doi.org/10.54324/j.mtl.v2i3.108

Kondo N. What Has Made Japan Healthy?: -Contributions of local and governmental health policies-. Japan Med Assoc J. 2014 Feb 1;57(1):24-7. PMID: 25237273; PMCID: PMC4130084.

Emily Washburn, “Daily Step Targets For Americans Are Easier Than Ever, So Why Are They So Hard To Hit?” Forbes, 2023 https://www.forbes.com/sites/emilywashburn/2023/03/28/daily-step-targets-for-americans-are-easier-than-ever-so-why-are-they-so-hard-to-hit/

Hirokage Tabata, “Walking 8,000 steps even 1-2 days a week reduces mortality rate: Japan study,” The Mainichi, 2023 https://mainichi.jp/english/articles/20230330/p2a/00m/0sc/012000c

Althoff, Tim et al. “Large-scale physical activity data reveal worldwide activity inequality.” *Nature* vol. 547,7663 (2017): 336-339. doi:10.1038/nature23018

Ross C. Brownson, Elizabeth A. Baker, Robyn A. Housemann, Laura K. Brennan, and Stephen J. Bacak: Environmental and Policy Determinants of Physical Activity in the United States, American Journal of Public Health **91**, 1995_2003, https://doi.org/10.2105/AJPH.91.12.1995

Ikejima, Yoshifumi. (2015). The Reality of Food Deserts in a Large Japanese City and Their Resolution Using Urban Agriculture. 10.1007/978-981-287-417-7_18.

Keyhani S, Wang S, Hebert P, Carpenter D, Anderson G. US pharmaceutical innovation in an international context. Am J Public Health. 2010 Jun;100(6):1075-80. doi: 10.2105/AJPH.2009.178491. Epub 2010 Apr 19. PMID: 20403883; PMCID: PMC2866602.

Buggle, J.C. Growing collectivism: irrigation, group conformity and technological divergence. J Econ Growth 25, 147–193 (2020). https://doi.org/10.1007/s10887-020-09178-3

Mark Zachary Taylor, Sean Wilson, Does culture still matter?: The effects of individualism on national innovation rates, Journal of Business Venturing, Volume 27, Issue 2, 2012, Pages 234-247, ISSN 0883-9026, https://doi.org/10.1016/j.jbusvent.2010.10.001.

Katherine Keisler-Starkey et. al, United States Census Bureau “Health Insurance Coverage in the United States: 2023,” 2024 https://www.census.gov/library/publications/2024/demo/p60-284.html#:~:text=In%202023%2C%20most%20people%2C%2092.0,percent%20and%2036.3%20percent%2C%20respectively Accessed January 2025

Tseng CW, Masuda C, Chen R, Hartung DM. Impact of Higher Insulin Prices on Out-of-Pocket Costs in Medicare Part D. *Diabetes Care*. 2020;43(4):e50-e51. doi:10.2337/dc19-1294

Oladimeji Ewumi, MedCentral, “The High Costs of Insulin,” https://www.medcentral.com/endocrinology/diabetes/the-high-costs-of-insulin Accessed January 2025

Yasuo Masai, “People of Japan,” Britannica, updated 2024 https://www.britannica.com/place/Japan/People

Chihiro Ara, The Asahi Shimbun, “Study: Japan will need 4 times as many foreign workers by 2040” https://www.asahi.com/ajw/articles/14540272 Accessed January 2025

Michael L. Anderson et. al, Brookings, “How the quality of school lunch affects students’ academic performance,” 2017 https://www.brookings.edu/articles/how-the-quality-of-school-lunch-affects-students-academic-performance/ Accessed January 2025

Persuasive Marketing and Manipulation Tactics: Exploring the Impact of Influencers on Purchase Decisions of Luxury Consumers

By Alina Yu (student)[1], Dr George Zifkos (mentor)[2*]

*1. Ascham School, Sydney, Australia
2. University of Leeds, UK

Abstract

As contemporary communication becomes increasingly reliant on social networks, the impact of Social Media Influencers (SMIs) on consumer purchasing behaviours similarly increases. SMIs can heavily affect consumer decisions through various manipulation tactics utilised in their online content, critically affecting the sales success of a luxury product. This study examines how effective the manipulation tactics which influencers use are upon Generation Z's purchasing decisions. Through a self-administered online survey, the study explored the impacts of SMIs on luxury consumers regarding their purchasing decisions as well as their perceptions of SMIs. By cross-tabulating the responses, the study indicated that although Gen Z luxury consumers were aware of how influential SMIs were, they still did not fully recognise the extent of the impact SMIs held on their luxury products purchasing decisions. Our findings also discovered a positive correlation between a consumer's online presence (followers and following count) and subjection to SMIs' influence.

Keywords

Behavioral and Social sciences; Sociology and Social Psychology; Social Media Influencers; Luxury goods; Manipulation tactics; Marketing

Introduction

Since the dawn of the 21st century, the domain of marketing has observed significant changes with social media emerging as one of the most influential marketing platforms that affect consumer purchasing decisions. The rise of Social Media Influencers (SMIs), individuals who create digital content and have a substantial number of followers, has created both opportunities and challenges for the luxury sector. This paper aims to explore the effects of marketing and manipulation tactics employed by SMIs on the purchasing decisions of Generation Z luxury consumers, "the generation of people born between 1997 and 2012" (Library of Congress, 2023).

Luxury products provide consumers with feelings of high exclusivity and status. The sales of these products rely on the perception of their qualities which can be greatly enhanced through SMI endorsements. SMIs substantially sway the public's perception of luxury products by creating particular relationships with consumers and therefore have the potential to impact their purchasing behaviour. This can effectively boost sales of luxury products, or drastically impact them. As these SMIs hold a considerable amount of influence over their audience, they also often create false lifestyles and adopt deceptive marketing to exercise personal goals such as profit.

The paper focuses on the impact of SMIs on younger generations (adolescents) and investigates how the latter are heavily influenced by online platforms. It will explore the strategies which SMIs employ to promote luxury goods/services and provide a more nuanced understanding of this contemporary phenomenon. It will uncover emerging trends that may shape the future of luxury marketing and the determinants of purchase decisions of luxury consumers.

It will first explore contemporary literature on how social media influencers utilise persuasive marketing and manipulation tactics to potentially affect consumers' purchasing decisions for luxury products. It will draw from a wide range of literature from the domains of consumer psychology, marketing, social media, and fashion research. Through an analysis of an online survey, it will provide a better understanding of the potential of social media marketing for luxury brands. This paper will also provide a better understanding of the dynamics between SMIs, luxury consumers, and luxury brands in the digital age.

The Rise of Social Media Influencers

The field of social media influencers is dynamic and continually evolving, with new platforms emerging and changes in user behaviour. Social media influencers are individuals who leverage social media platforms to build and engage with a dedicated audience. They have the ability to shape the opinions, preferences, and behaviours of their followers, making them influential in specific niches or industries. Academic discussions on social media influencers often focus on their role in digital marketing, brand partnerships, and the impact they have on consumer behaviour.

Influencers range from all types of people making all types of content. Some make content purely for fun while others make it their career: they make an income by taking on advertisements, giving reviews, or simply aiming for enough interaction for the platform itself to pay them. Ouvrein et al. (2021) argued there were three types of influencers: "Passionate Business Influencers, Passionate Influencers and Celebrity Influencers" (p.1313). Passionate business influencers exploit social media to generate income either as a full-time or part-time occupation; passionate influencers use social media to share a desired motive; and celebrity influencers utilise social media to gain publicity. There is no universal definition for SMIs nor a clear distinguishing trait differing the types of influencers, and is an extremely flexible term.

Social media influencers start out as regular people posting content online and achieve success from high amounts of interaction. Since they do not have to be qualified and their industry is not regulated, many people, especially the younger generation, attempt to post content in an effort to go 'viral' – "quickly and widely spread or popularized especially by means of social media" (Merriam-Webster Dictionary, 2024). The fact that many influencers appear to be 'normal', everyday people makes them seem more trustworthy and relatable in their content, allowing them to promote products/services in a discreet manner, though they must be careful with the products/services they promote. Since influencers are generally perceived as more reliable and trustworthy, they affect consumers' purchasing decisions more than any other type of marketing. Schouten et al. (2021) stated that consumers find influencers more similar to

themselves than celebrities however consumers would rather be more similar to celebrities than influencers, therefore the consumer's purchase decision depends on how similar they find themselves with the endorser and how much they want to be like them.

However, an influencer's reputation is extremely versatile and will change at any given moment depending on their actions, their reviews, what they post, who they post and what other influencers say about them. This means that influencers often have a very short-lived career and are very prone to 'cancel culture' which means "they will lose their audience, most probably brand deals, and their fame" because the public disliked their actions (Gündüz, 2021). Feng et al. (2023) stated that influencers balance a thin line between generating trust or envy and depending on very subtle details in their communication method, influencers can easily blur the lines between personal and professional life, generating both online and in-person hate.

Mains (2023) said that some key characteristics of a successful influencer consist of authenticity, relatability, engaging content… and that they are "effective in promoting brands, products, or services". To be as influential as possible, SMIs try to expand their platform without losing previous engagers. To do so, they employ various tactics to encourage loyalty and long-term relevance, such as recreating their most popular content.

Manipulation Tactics

SMIs hold great power to influence and mobilise their followers into making decisions regarding their lifestyle and purchases. Like celebrities, many SMIs are perceived as trusted sources of information for millions of followers. Nevertheless, amidst their glamorous lives, there lies a dark reality: the purposeful use of manipulation tactics to influence opinions, behaviours, and purchasing decisions. The arsenal of manipulation tactics employed by social media influencers is vast and multifaceted. Emotional manipulation lies at the heart of many influencer strategies, fostering fear of missing out (FoMO), exploiting insecurities, leveraging emotional triggers for increased engagement and social engineering, and practising deceptive marketing.

SMIs often use visually appealing content and relatable narratives to establish a personal connection with their followers, fostering a sense of trust and admiration. For instance, major TikTok users as: @alixearle, @tarayummy, and @hayleyybaylee leverage authenticity, humour and relatability to connect with their audience at a personal level and create a relatable or desirable image. @hayleyybaylee (Haley Kalil) for example, claims to be "not elite" and just "a normal person" despite having more than 10 million followers and having accumulated tremendous levels of wealth as an American model/SMI (Clark, 2024).

Through meticulously curated images and lifestyles, influencers create aspirational narratives that position luxury products as 'must-have' possessions, driving desire among their audience. However, in reality, such manipulation tactics are not as effective when exposed and can often result in negative responses from both the SMI and the brand's image. Consumers are becoming "increasingly aware of influencers' affiliation to brands, as evidenced by recent industry reports" (Statista, 2018).

The FoMO Phenomenon

The FoMO (Fear of Missing Out) phenomenon, a term coined in the early 2000s, refers to the apprehension that others are experiencing rewarding experiences from which one is absent. It is a widely used term today to describe “the perception of missing out” and “is closely related to the fear of social exclusion or ostracism, which existed long before social media” according to a Harvard psychology professor (Dattilo, cited in Laurence, 2023). FoMO exerts a profound influence on individual behaviour, driving engagement with social media platforms and consumption of influencer-generated content. The fear of missing out on the latest trends, events, or experiences compels individuals to continuously monitor their social media feeds to seek validation and reassurance through post engagements as: likes, comments, and shares. Moreover, FoMO prompts impulsive behaviour, leading individuals to make purchasing decisions or attend events solely to align with the perceived lifestyles of influencers. Despite the phenomenon dating to pre-social-media (1997), such as purchasing the latest technology (flip phones in 1996), the term FoMO itself originates from social media. The term was introduced in 2004 and was only widely used in 2010 to “describe a phenomenon observed on social networking sites" (Gupta and Sharma, 2021). Influencers use various persuasive tactics to engage and captivate their audience to promote certain products/services, often taking advantage of the FoMO effect to pressure viewers into purchasing luxury goods and services.

A Harvard psychology professor (Dattilo, cited in Laurence and Temple, 2023) also said that FoMO “triggers anxiety, and compulsive behaviours, like checking and refreshing sites, to maintain social connections” Research has linked excessive social media use and FoMO to heightened levels of anxiety and diminished psychological well-being, highlighting the need for critical reflection and moderation in consumption habits. The incessant exposure to curated depictions of perfection can have detrimental effects on both the consumer’s and influencer’s mental health. Comparing one's own life to the seemingly flawless portrayals of influencers can breed feelings of inadequacy, low self-esteem, and depression. Social media influencers often perpetuate the illusion of authenticity, blurring the lines between reality and fiction. Behind the meticulously staged photos and polished captions lies a carefully curated narrative designed to appeal to followers and sponsors alike. This veneer of authenticity can obscure the realities of the influencer's life, fostering unrealistic expectations and perpetuating unattainable standards of success and happiness.

The Bandwagon Effect

The bandwagon effect is a powerful force that drives individuals to conform to prevailing trends and adopt popular opinions. It closely relates to Festinger’s (1954) Social Comparison Theory, defined as “the idea that people evaluate their own opinions, values, achievements, and abilities by comparison respectively with the opinions, values, achievements, and abilities of others” (Powdthavee, 2014). Similar to the FoMO phenomenon, influencers leverage it to sway public perceptions and foster a culture of conformity in the digital realm. SMIs, especially those more popular, capitalise on this psychological phenomenon of consumers by showcasing and

exploiting their immense audience size while exaggerating their extravagant lifestyle and fame. Many studies describe this phenomenon as 'cognitive bias' wherein individuals are inclined to adopt beliefs or behaviours simply because they perceive them to be popular or widely accepted. By creating an illusion of unanimity and social approval, SMIs compel their followers to jump on the bandwagon and emulate their actions, whether it be purchasing a particular product, endorsing a specific ideology, or adhering to a certain lifestyle. The bandwagon effect further manifests itself by creating a domino effect to influence more people to 'jump on the bandwagon' as there are more people involved in the trend. This cultivates exclusivity, further increasing the demand and portraying the popularity of the product; this is often seen in businesses deliberately producing slightly less than the demand so when stocks are sold out, consumers view it as exclusive and further generate demand. For example: Erewhon, an exclusive and luxurious supermarket with increasing popularity due to its image on social media is a result of SMIs promotive content, especially social trends stemming from the well-known 'trend-setting' celebrity Hailey Bieber and her famous "Hailey Bieber strawberry smoothie" costing $19 USD. As a result, followers are more likely to emulate Hailey's style and purchase the same items, perpetuating the cycle of conformity and reinforcing the bandwagon effect. As Erewhon "seems tailor-made for today's influencer culture" (Berlinger, 2021), it further proves how the bandwagon effect manifests itself as many of its consumers are also influencers who create content to further influence others.

Therefore, influencers strategically align themselves with particular lifestyles and position themselves as "trend-setters", especially celebrity influencers. SMIs specifically leverage their specialised platform to portray the desired effect whether genuine or not, some SMIs may also purchase followers to enhance their portrayed popularity. Specific terminology and words are also used to bait consumers into believing the popularity of a product/lifestyle for eg: "everyone" and "trending"; such terms trigger responses from consumers to an increased desire to also be a part of a social trend, or purchase something 'everyone else' has.

Deceptive Marketing

SMIs also often use deceptive techniques to achieve their promotional goals. This includes but is not limited to: undisclosed/hidden sponsorships, insufficient information whether deliberate or accidental, false reviews and claims. For example, not writing an advertisement disclaimer in the caption of their content, not providing enough details about the product's side effects, and lying about a product are all instances of deceptive marketing. This is often a last resort which SMIs use as it is extremely harmful to their reputation and could potentially dismantle and discredit all previous efforts to influence whether honest or not. Mikayla Nogueira, a major beauty influencer, was previously exposed for supposedly providing her audience with a false review on a mascara product and immediately received online criticism that questioned the authenticity of her content (Mendes II, 2023). Her mascara review was also a paid partnership however, #Ad was not in her caption which TikTok requires branded content to have (Weir, 2023). Regardless of whether Mikayla had fraudulent intentions or whether she was

misunderstood, it can be seen that if viewers detect or determine any deceptive techniques being used, the SMI will face detrimental damages to their reputation.

Deceptive marketing also includes the purchasing of false engagement to one's social media profile. By buying 'bot' followers, which are "automated or semi-automated accounts" (Brisset, 2024), SMI's will suffer credibility loss. As an SMI's followers count heavily influences viewers' perception of them, buying them also means deceiving viewer's of their reputation and reliability. A high number of followers helps boost the SMI's brand reputation and increase trust from viewers as a reliable and reputable source ('Instagram follower count', 2024). Fake followers on the other hand, signal to the SMI's audience that they are not as reliable as they portray themselves to be on their social media page and must resort to deceitful tactics to improve brand reputation and credibility.

The Emotional Appeal of Luxury and Lifestyle Integration

Luxury brands leverage social media influencers to seamlessly integrate their products into the influencer's lifestyle. By showcasing these products in relatable, everyday situations, influencers make luxury items appear attainable and integral to an aspirational lifestyle. The emotional appeal of such content taps into the audience's/followers' desires for status, sophistication, and exclusivity, fostering a strong emotional connection that influences purchasing decisions.

In 2022, the market for luxury goods was valued to be worth USD 272.74 billion worldwide and is expected to rise to USD 392.40 billion by 2030 ('Luxury Goods Market Size', 2024). People purchase luxury products daily because of their popularity in modern society. Besides fundamental products needed to stay alive, the majority of the population in a first world country will purchase luxury products and services to improve their quality of life, especially with the rise of e-commerce. Since globalisation, the use of technology has vastly increased. This has allowed purchases to be made all over the world much more quickly and more straightforwardly. Consumers will be able to purchase globally for various reasons, but all consumptions are universally bought to enhance one's emotional state to their will.

Luxury products and services do not carry an objective definition but multiple scholars have attempted to define them. Luxury products range from non-essential goods but can be categorised into eight groups which Chevalier and Mazzalovo (2008) identified as "fashion, jewellery, cosmetics, wine, automobile, hotel, tourism, and private banking". Kenton (2024) defines luxury goods as "not necessary to live, but it is deemed highly desirable within a culture or society.", similar to the Cambridge Dictionary's (2024) definition of luxury goods being "expensive things, such as jewellery and make-up, that are pleasant to have but are not necessary". Eckman et al. (1990) recognized luxury products to have "high aesthetic and hedonic values". Consumers view fashion merchandise to have high social and symbolic value (Bilro, 2021).

While the list of luxury goods is not defined explicitly, the universal understanding of luxury goods is that they are not a necessity but solely for pleasure. Purchases of luxury products

also depend on the consumers' evaluations of this product/service. Due to limitations such as income, the consumer may choose not to purchase certain luxury products and purchase an alternative good based on their perception of the good's value. This is because luxury goods "tend to be sensitive to a person's income or wealth" (Kenton, 2024). Therefore, they are commonly associated with wealth and status, and consumers who wish to appear wealthy or simply wish to enjoy their wealth due will often purchase more luxury goods than those with lower incomes. Such purchasing behaviour of non-essential items has evolved to become the norm in modern society, especially in first-world countries. Businesses respond to these consumers by using psychological manipulation tactics to generate sales, such as FoMO. They use the idea of pleasure and satisfaction to persuade consumers into believing certain luxury items are a necessity. Following globalisation, businesses have also changed advertisement methods for luxury goods, shifting from traditional methods such as runway models and magazines to social media interactions. Nike, for example, has been able to utilise social media greatly as "90% of [consumers] buy from brands they follow on social media" (Williams, 2020). Nike uses athletes as SMIs to increase brand awareness and online presence, leveraging the athlete's existing reputation and followers. By doing so, Nike has been able to amass over 305 Million followers on Instagram as of July 2024 and grew its global annual sales by 10 per cent to $51.2 billion (McGee 2023).

Vinerean and Opreana state that "luxury fashion brands have the potential to induce psychological and emotional values among customers and aspirational consumers especially because these brands have prestige, price, craftmanship, exclusivity, and in certain cases, heritage." (2019). Hérmes is a prime example of a luxury brand that is extremely successful in their marketing strategy as an extremely exclusive brand. By creating a perception amongst consumers of scarcity and exceptional quality through their limited production, Hérmes creates a surplus of demand from their famous Birkins and Kellys which flows into other products such as belts (Roll, 2020).

Methodology

SMIs are able to impact the purchase decisions and consumption of luxury goods from consumers through various marketing and manipulation tactics. The main objective of this study is to investigate what factors influence consumer purchasing decisions and how SMIs can also impact them. The study focuses on Gen Z consumers as they are the most impacted age group of social media. Therefore, this study explores who and how social media influencers on purchase decisions of luxury goods and services.

Since the target population of this study was Gen Z consumers, an online questionnaire was designed and distributed via social media. This study recruited 65 respondents who were approached via Twitter, online communities, and word of mouth. This survey was available to participants for seven weeks. The invitation included a link to the online survey and information about the purpose of the study, the voluntary nature of participation, and assurances of confidentiality. Since we believe that not all participants share the same understanding of luxury

goods and services, we also included Ko et al.'s (2019) template definition in the survey instrument. According to Ko et al., all luxury purchases must:

> 1) be high quality;
> 2) offer authentic value via desired benefits, whether functional or emotional;
> 3) have a prestigious image within the market built on qualities such as artisanship, craftsmanship, or service quality;
> 4) be worthy of commanding a premium price; and
> 5) be capable of inspiring a deep connection, or resonance, with the consumer. (p.2, italics in original)

The survey was developed using Google Forms. The survey consisted of 21 items, divided into three sections: demographic information, questions designed to assess participants' perceptions of SMIs and their impact, and questions attempting to measure key variables of interest such as social media usage and online presence. The first section of the survey consisted of multiple-choice questions for the participants to answer, collecting demographic information. Multiple-choice questions categorise the participants into groups that are more easily analysed. By comparing certain demographics, repeated patterns regarding social media usage and its impact can be established. Multiple-choice answers create a simple, quantified response from the participants and clearly shows their overall circumstances regarding social media and lifestyle. Next, the survey consisted of questions measured over Likert scales and multiple-choice questions for participants to rate the impact of SMIs on both themselves and those around them. Likert scale questions were used in the second section to capture participants' perception of SMIs and their impact, with one (1) being the least impactful, to ten (10) being most impactful. By using Likert scales, the participant's perception of SMIs and social media is also quantified, showing distinct trends if available. This section attempted to elicit the attributes of participants' purchasing decisions as well as aspects of their interaction with social media. Lastly, multiple-choice questions were used, again, to collect the participants' social media profiles. The last section of the survey was designed to determine the participant's social media profile and usage. Exploring their social media profile and presence is crucial as it directly relates to their exposure to SMIs, impacting their perception of SMIs and their influence.

The results of the survey were analysed using Google Forms and Excel to run frequencies and cross-tabulations. The study aimed to explore whether particular demographic variables have an impact on participants' perceptions of SMI and/or their social media life. The survey also intended to explore whether particular social media usage patterns have an effect on an individual's purchasing behaviour.

As there were only 65 participants and the majority of participants were based in Australia, it may not be an accurate representation of the impact of SMIs on all Gen Zs. The essence of distributing the survey online in various social media platforms and online communities may have skewed the results as all participants would have access to technology

already. Participants from outside Australia may have also skewed the results in terms of spending due to SMIs as they may have entered their amounts without converting to AUD.

Results

92.3% of respondents were between ages 10-18, 6.2% were ages 19-25, and there was one outlier who was between ages 36-45. 73.8% of responders identified as female and 21.5% were male. The current employment statuses of the respondents consisted of 86.2% students. The respondent's demographic profiles are shown in Table 1:

Demographic	Frequency	Percentage (%)
Gender		
Female	48	73.8
Male	14	21.5
Other	3	4.7
Age		
10-18	60	92.3
19 - 25	4	6.2
26- 35	0	0
36 - 45	1	1.5
Current employment status		
Student	56	86.2
Unemployed	2	3.1
Employed full-time	1	1.5
Employed part-time	4	6.2
Self-employed	2	3.1
Annual income ($ AUD)		
Student	40	61.5
0 - 20 000	22	33.8
20 001 - 45 000	1	1.5
45 001 - 100 000	1	1.5
100 001+	0	0
Prefer not to say	1	1.5
Level of completed education		
Primary (year 1 - 6)	6	9.2
Secondary (year 7 - 9)	32	49.2
Upper secondary education (year 10 - 12)	25	38.5
university degree	1	1.5
Post-graduate	1	0
PhD or higher	0	0
None	0	0

Weekly working hours (including education):		
0 - 5	5	7.7
6 - 15	12	12.3
16 - 25	3	4.6
26 - 35	15	23.1
36 - 40	20	30.8
40+	14	21.5

Table 1: Respondents demographic profiles

To analyse the data, comparisons between the respondents' socioeconomic characteristics, views on SMIs, and purchasing decisions for luxury goods are made. With such comparisons, correlational analysis is also performed to make out any patterns that may influence the participant's responses.

Table 2: Likert data

Constructs	**Min.**	**Max.**	**Mean**	**Median**	**Std. Deviation**
Likelihood of being influenced by SMI	1	9	4.52	5	2.37
Likelihood for others to be influenced by SMI	1	10	6.29	7	2.40
Trust towards SMI	1	8	3.95	4	1.79
SMI relatability	1	9	4.58	5	2.05
Trust towards social media	1	9	4.25	4	1.98

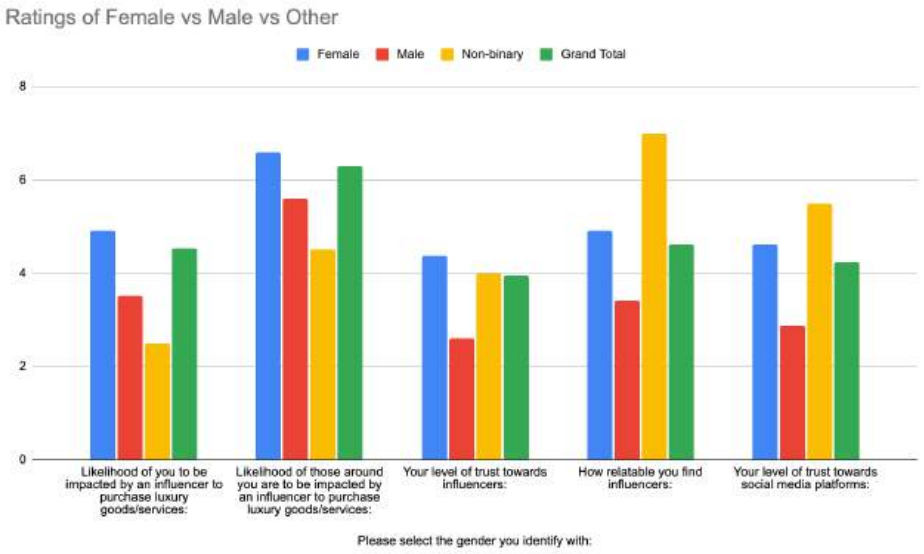


Figure 1: Gender-based responses

In Question 2, responders were asked to enter their associated gender. Answers from Questions 7, 8 14,15,16 respectively were then cross-tabulated with their corresponding gender. By comparing the responses between females, males, and others, a distinct correlation is found between all ratings regarding impact from SMIs and trust towards them. Females have consistently responded with higher ratings regarding their likelihood and those around them's likelihood of being influenced by SMIs regarding their purchasing decisions. They have also consistently rated their level of trust and relatability towards SMIs and social media platforms higher than male respondents. Responses from 'other' genders may not be accurate, however, due to their small sample of respondents.

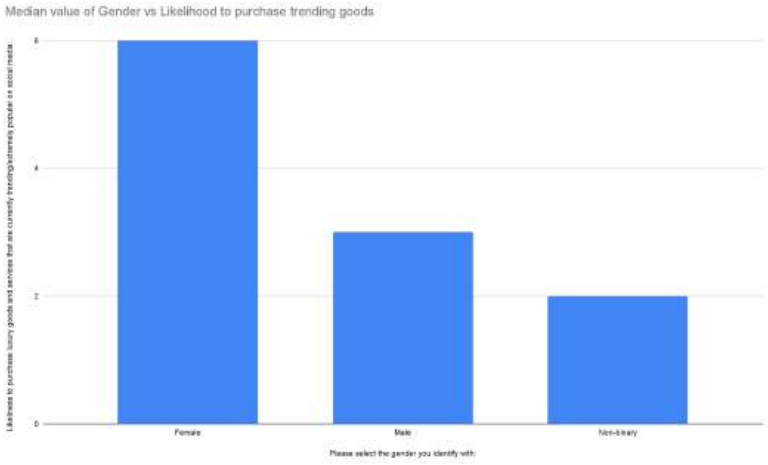


Figure 2: Gender-based responses (Q9)

The median rating between 1-10 (10 being mostly likely) for females to purchase currently trending goods was 6, while the median male rating was 3 and non-binary respondents rated 2.

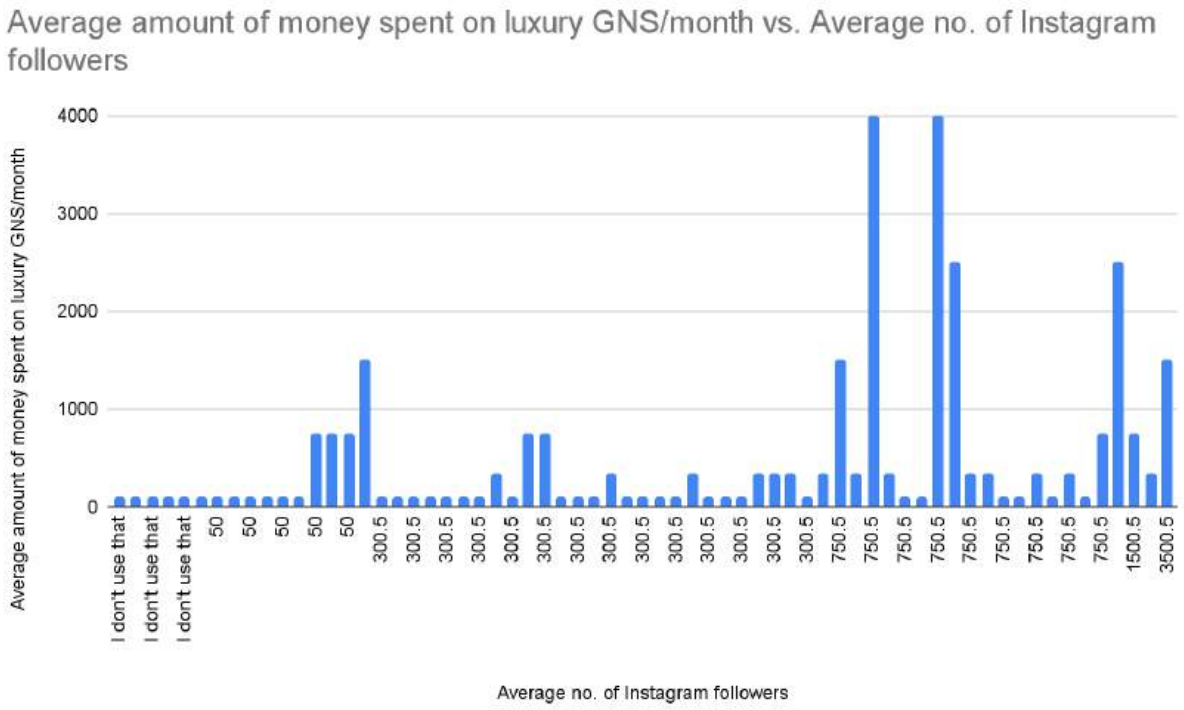


Figure 3: Social Media Profile based responses

By looking at the profile size of each respondent via. their number of Instagram followers (Q19) and comparing their average monthly spending on luxury goods (Q12), a correlation can be seen where those with larger profiles (more Instagram followers), spend more monetarily on luxury goods. Those who do not use Instagram have all reported spending $0-100 AUD on average per month while, on the other end, those with greater than 700 Instagram followers have reported spending up to $4000 AUD on luxury goods per month. Overall, profiles with 500-1000 Instagram followers spend the most, and the fewer respondents with followers between 1000-2000 and 2000-5000 still spend a vast amount compared to those with under 500 Instagram followers.

Figure 4: How responders view SMI impact on themselves and those around them

Responders have consistently interpreted the power of SMIs to influence their audience to purchase luxury goods to be greater on those around them (Q8) than themselves (Q9). In almost all responses, respondents have rated SMIs' influence on those around them to be equal or greater than their own. The median rating of how likely those around the respondents are to be impacted by SMIs is 7 out of 10, 10 being the most likely to be impacted. On the other hand, the median rating of SMI influence on themself is only 5.

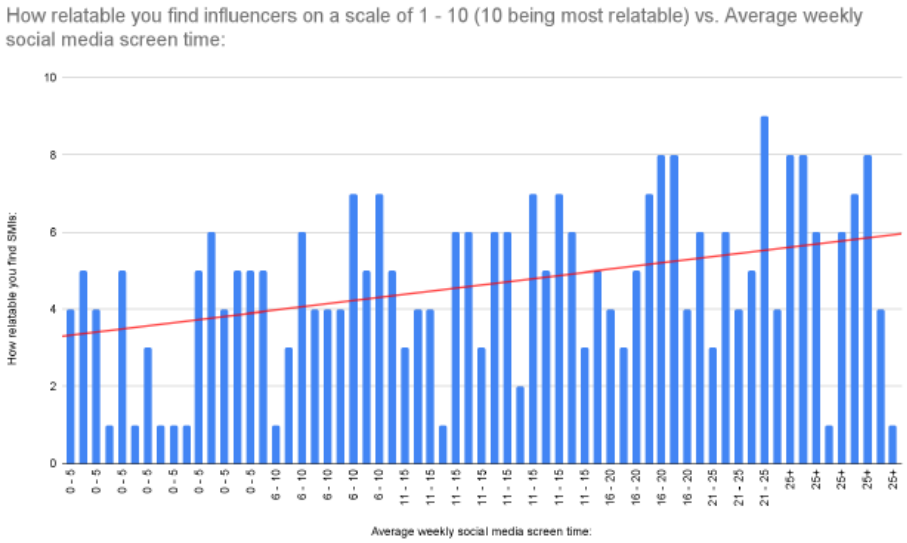


Figure 5: Screen time vs perceived relatability with SMIs

By comparing the average screen time of responders (Q17) and how relatable responders find SMIs to be (Q15), a positive correlation was found where those with larger screen times generally found SMIs more relatable.

Discussion

From the online survey consisting 98.5% of 65 participants between ages 10-25, it was assessed that while the survey's participants acknowledged the strong influence SMIs hold upon their peers, they perceive themselves to become less susceptible to the same influence (Q7,8). The majority of answers settled towards 8 or 9 out of 10 (10 being most impacted) on the Likert scale regarding SMI's influence on those around them, while the average rating for SMI's impact on themselves was 5. Over 95% of participants rated SMI influence over those around them higher than themselves. This paradoxical relationship and cognitive dissonance suggest a potential defence mechanism for consumers as they are trying to protect their autonomy when it comes to purchasing decisions. This perception that others are more easily influenced may reflect societal pressures and norms which the research participants may or may not be conscious of.

Gender differences in trust have also emerged as a significant finding in this study. Female respondents rated their level of trust towards SMIs higher in comparison to the male counterparts of this study (Q7). This indicates a greater level of susceptibility amongst females to SMIs compared to the male participants of this study. Bizarrely, the female participants acknowledged their greater level of susceptibility to SMIs. Similarly, female participants rated the influence of SMIs over those around them higher than male participants, acknowledging the influence SMIs have over consumer purchasing decisions. Such gender discrepancies may be attributed to the types of products which influencers endorse, which mainly consist of fashion and beauty products (Q10,11).

The relationship between one's social media presence and relatability with SMIs has also generated a notable insight in this study. Respondents with greater online presence report increased relatability and trust towards SMIs and are much more susceptible to purchasing luxury goods. Over 40% of participants expressed a likelihood of purchasing trending luxury items due to the social media exposure, which has been created. This indicates an underlying influence that consumers may not have fully acknowledged in this study.

The relationship between screen time (Q17) and perceived relatability with SMIs (Q15) adds another depth to this discussion. Individuals with longer screen times, on average, have also reported increased feelings of connection with SMIs compared to participants with lower than average screen times. This suggests that increased exposure to social media may enhance feelings of relatability and trust with SMIs. This finding shows that increased time spent on social media can influence consumer perceptions and behaviours, and will therefore cultivate increased genuine relationships between consumers and SMIs. Such relationships can enhance the effectiveness of social media marketing, allowing SMIs to be more effective in influencing consumer purchasing decisions.

Conclusion

By examining responses from individuals ranging from ages 10-25 and a singular participant within the age range of 36-45, the study was able to identify various consistent effects of luxury consumers due to SMIs. The study was able to draw out the effectiveness of SMIs in influencing consumer purchasing decisions, regardless of whether participants had realised or not. The employment of various digital manipulation strategies by SMIs have evidently been effective in influencing what their audience purchases, and the amount they purchase. Although most consumers are mostly unaware of their susceptibility to these influences, they still evidently are affected by SMIs. It's also been discovered that online presence also contributes greatly to how susceptible a consumer is, being that the more present one is online, the more susceptible they are to SMI manipulation tactics. Thus, consumers, more likely than not, fall prey to persuasive marketing tactics on social media platforms, especially if they are highly engaged with the platform themselves. Especially with a rise in social media popularity, SMIs' interaction with consumers increases which, therefore, enhances their effectiveness in influencing consumer purchasing decisions.

This study contributes to our understanding of how luxury consumers perceive SMIs, and how influential SMIs are in driving luxury consumption. While SMIs are extremely effective in manipulating their audience, ethical considerations remain as the autonomy of consumers may be undermined. Future research could explore the broader implications for marketing ethics in the digital space and the role of authenticity in shaping consumer trust. As this study only employed members of the Australian audience, further studies may be conducted on the cross-cultural differences of SMIs impacts and how luxury consumers respond to them.

Acknowledgements

I would like to thank my family and teachers for supporting my passions and my pursuit of knowledge outside the classroom. This paper would not have been realised without the insightful feedback and encouragement of my mentor Dr George Zifkos, assisting me with his valuable expertise. Lastly, I want to thank all my anonymous participants whose views have been instrumental in shaping this work. I am forever grateful for everyone's support and inspiration.

Works cited

Berlinger, M. How Erewhon Became L.A. 'S Hottest Hangout (Published 2021). The New York Times. 2024. https://www.nytimes.com/2021/02/17/style/erewhon-los-angeles-health-food.html (accessed 2024-05-02).

Bilro, R. G.; Maria, S.; Fonseca, J. Masstige Strategies on Social Media: The Influence on Sentiments and Attitude toward the Brand. International Journal of Consumer Studies **2021**, 46 (4), 1113–1126. https://doi.org/10.1111/ijcs.12747.

Brisset, C. Understanding the Influence of Social Media Bots | ICUC. ICUC. https://icuc.social/resources/blog/influence-of-social-media-bots/#:~:text=Social%20media%20 bots%20are%20 automated,their%20impact%20on%20user%20 experience. (accessed 2025-01-01).

Büşra Gündüz. Cancel Culture: A Harsher Criticism for the Famous. INFLOW Network. https://inflownetwork.com/cancel-culture-a-harsher-criticism-for-the-famous/ (accessed 2025-01-01).

Chevalier, M.; Mazzalovo, G. LUXURY BRAND MANAGEMENT : A New World of Privilege.; John Wiley, 2008.

Clark, M. Influencer Haley Kalil apologises for “let them eat cake” Met Gala video. The Independent. https://www.independent.co.uk/life-style/haley-kalil-met-gala-video-apology-b2545044.html (accessed 2025-01-01).

Eckman, M.; Lynn Damhorst, M.; J. Kadolph, S. Toward a Model of the In-Store Purchase Decision Process: Consumer Use of Criteria for Evaluating Women’s Apparel - Molly Eckman, Mary Lynn Damhorst, Sara J. Kadolph, 1990. Clothing and Textiles Research Journal **2016**, 8 (2).

Feng, W.; Chang, D.; Sun, H. The Impact of Social Media Influencers’ Bragging Language Styles on Consumers’ Attitudes toward Luxury Brands: The Dual Mediation of Envy and Trustworthiness. Frontiers in Psychology **2023**, 13. https://doi.org/10.3389/fpsyg.2022.1113655.

Gaëlle Ouvrein; Pabian, S.; Giles, D.; Liselot Hudders; Backer, C. D. The Web of Influencers. A Marketing-Audience Classification of (Potential) Social Media Influencers. Journal of Marketing Management **2021**, 37 (13-14), 1313–1342. https://doi.org/10.1080/0267257x.2021.1912142.

Gupta, M.; Sharma, A. Fear of Missing Out: A Brief Overview of Origin, Theoretical Underpinnings and Relationship with Mental Health. World journal of clinical cases **2021**, 9 (19), 4881–4889. https://doi.org/10.12998/wjcc.v9.i19.4881.

Influencer marketing: consumer attitude in the UK 2018 | Statista. Statista. https://www.statista.com/statistics/822171/influencer-marketing-consumer-attitude-in-the-uk/ (accessed 2024-06-20).

Instagram Follower Count: The Impact of Instagram Follower Count on Brand Reputation - FasterCapital. FasterCapital. https://fastercapital.com/content/Instagram-Follower-Count--The-Impact-of-Instagram-Follower-Count-on-Brand-Reputation.html#:~:text=%2D%20Trust%3A%20A%20high%20follower%20count,information%2C%20products%2C%20or%20services. (accessed 2024-07-17).

John Paul Mains. What are the Top 5 (now 6) Characteristics of Social Influencers? Click Laboratory. https://www.clicklaboratory.com/blog/five-characteristics-of-an-influencer/ (accessed 2024-04-25).

Kenton, W. What Is a Luxury Item (aka Luxury Good)? Definition and Examples. Investopedia. https://www.investopedia.com/terms/l/luxury-item.asp (accessed 2025-01-01).

Ko, E.; Costello, J. P.; Taylor, C. R. What Is a Luxury Brand? A New Definition and Review of Literature. Journal of Business Research **2019**, 99 (1), 405–413. https://doi.org/10.1016/j.jbusres.2017.08.023.

Laurence, E. The Psychology behind the Fear of Missing out (FOMO). Forbes. October 16, 2023. https://www.forbes.com/health/mind/the-psychology-behind-fomo/ (accessed 2024-06-20).

Luxury Goods Market Size, Growth, Trends | Overview [2030]. Fortunebusinessinsights.com. https://www.fortunebusinessinsights.com/luxury-goods-market-103866 (accessed 2025-01-01).

McGee, B. Nike's Annual Revenues Exceed $51.2 Billion - Footwear Insight. Formula4media.com. https://www.formula4media.com/articles/nikes-annual-revenues-exceed-51-2-billion (accessed 2025-01-01).

Merriam-Webster Dictionary. Merriam-webster.com. https://www.merriam-webster.com/dictionary/viral#:~:text=%3A%20quickly%20and%20widely%20spread%20or%20popularized%20especially%20by%20means%20of%20social%20media (accessed 2024-12-26).

Moises Mendez II. A Mascara Drama Unfolding on TikTok Raises Questions About Authenticity Among Beauty Influencers. TIME. https://time.com/6250881/mikayla-nogueira-mascara-fake-eyelashes/ (accessed 2024-06-20).

Nattavudh Powdthavee. Social Comparison Theory. Springer eBooks **2014**, 6028–6029. https://doi.org/10.1007/978-94-007-0753-5_2740.

Research Guides: Doing Consumer Research: A Resource Guide: Generations. Loc.gov. https://guides.loc.gov/consumer-research/market-segments/generations#:~:text=Generation%20Z%2C%20 also%20 sometimes%20 known,a%20part%20of%20the%20 lives. (accessed 2024-12-26).

Roll, M. Hermès - The Strategy Insights Behind The Iconic Luxury Brand – Martin Roll. Martin Roll.

https://martinroll.com/resources/articles/strategy/hermes-the-strategy-behind-the-global-luxury-success/ (accessed 2025-01-01).
Schouten, A. P.; Janssen, L.; Verspaget, M. Celebrity vs. Influencer Endorsements in Advertising: The Role of Identification, Credibility, and Product-Endorser Fit. International Journal of Advertising **2019**, 39 (2), 258–281. https://doi.org/10.1080/02650487.2019.1634898.
Vinerean, S.; Opreana, A. Social Media Marketing Efforts of Luxury Brands on Instagram. Expert Journal of Marketing **2019**, 7 (2), 144–152.
Weir, G. TikToker who wore false lashes in sponsored mascara review labelled "dishonest." Nine.com.au. https://style.nine.com.au/beauty/mikayla-nogueira-called-out-for-dishonest-mascara-review/2d4bedb6-3186-4db0-a04c-ea55c2567a26#:~:text=%22It%20 truly%20did%20a%20 phenomenal,wrote%20and%20the%20 resurfaced%20clip. (accessed 2024-07-17).
Williams, R. 90% of people buy from brands they follow on social media. Retail Dive. https://www.retaildive.com/news/90-of-people-buy-from-brands-they-follow-on-social-media/577399/ (accessed 2025-01-01).

Navigating the Ethics of Artificial Intelligence: A Literature Review By Jillina Weng

The recent soar in the development of artificial intelligence (AI) has taken the world by storm. From healthcare and education to autonomous vehicles and social media algorithms, AI's potential to revolutionize the way we live and work is unparalleled. Yet, despite these advancements, AI also raises critical concerns that extend beyond its technical capabilities: namely, how does the integration of AI into daily human activities engender ethical dilemmas? A review of the available literature indicates that AI presents profound ethical challenges by perpetuating societal biases through algorithmic decision-making, undermining privacy through extensive data collection, and reducing autonomy by creating overreliance on automated systems.

Some sources argue that AI reinforces existing societal biases due to its reliance on flawed or biased training data, leading to unfair outcomes. In a research report by Carina Prunkl, she explains that AI analyzes large amounts of data "on people's online profiles and online behaviour to identify statistical correlations", creates a model using this data, and then implements it when making new decisions (Prunkl 4). In an article by UNESCO, they highlight an example of gender biases prevalent in society: by typing "greatest leaders of all time" into a search engine, a list of prominent male figures will most likely show up. Women do not show up as often in this list, demonstrating the stereotypes "deeply rooted in our societies" (UNESCO). Christina Pazzanese, writing for the Harvard Gazette, agrees with this notion that there are many prejudices in society that are fed to AI systems and will thus be replicated by AI. She quotes political philosopher Michael Sandel, who argues that AI algorithmic decision-making appeals to society because of its objectivity that can possibly overcome human subjectivity and bias. Additionally, according to Joseph Fuller, professor of management practice at Harvard Business School, when used carefully and thoughtfully, AI software "allows a wider pool of applicants to be considered than could be done otherwise, and [could] minimize the potential for favoritism that comes with human gatekeepers" (Pazzanese). However, Michael Sandel contends that "many of the algorithms that decide who should get parole, for example, replicate and embed the biases that already exist in our society" (Pazzanese). She mentions that Sandel also states how AI possessing these prejudices and biases can confer a "scientific credibility" on them as if it has an objective status (Pazzanese). This is dangerous because the stereotypical representations that exist within society are based on subjective judgments made based on mere opinion, not fact.

In addition to biases and prejudice, AI systems compromise privacy by collecting and analyzing vast amounts of personal data, often without full user awareness and at times, without consent. Maria Stefania Cataleta from University Côte d'Azur writes in her research report that many governments are implementing AI and new technologies, such as video surveillance and biometric tracking, to prevent illegal activity and make lives more secure. At the same time, she states that these technologies "actively monitor and track common citizens, which constitutes a violation of individual privacy" (Cataleta 1). And while we are generally informed of when our data is being used and asked for consent, it can also be "fraudulently extracted and used by

criminal networks to extort for money" (Cataleta 3). Just by being online and partaking in any form of digital communication, we put our privacy at risk. According to Caitlin Chin-Rothmann, writing for the Center for Strategic and International Studies (CSIS), private information can be revealed even if anonymity is maintained. She explains that "algorithms can profile individuals based on otherwise unconnected data points", thus piecing together and revealing private details through these outcomes (Chin-Rothmann). For example, by analysing shopping history, internet browsing activity, and geolocation, AI systems can deduce details about a person's income, religion, political stance, and more (Chin-Rothmann). In workplaces as well, employee data and privacy concerns are constantly held at debate. In a Harvard Business School blog post by Kate Gibson, she asserts that "AI systems can analyze vast amounts of personal and professional information, which must be properly protected to avoid privacy violations, unauthorized access, and misuse" (Gibson). Additionally, with AI systems handling and storing sensitive data, cybersecurity concerns are rising and these databases are more prone to cyberattacks (Gibson). AI can facilitate more targeted phishing attacks or other scams based on personal information. Chin-Rothmann illustrates how attackers can "impersonate victims using synthetic media or tailor deceptive messaging to specific people" (Chin-Rothmann). Indeed, algorithmic privacy violations come with concrete economic, security, and reputational harms to our society.

Lastly, delegating ethical decisions to AI in critical situations risks oversimplifying complex human values. As AI takes over so many tasks that were previously performed by humans, the role of subjective, human judgement in decision-making becomes less clear. Sandel mentions that, while concerns over bias and privacy in AI are familiar by now, we still need to consider if AI and smart machines can outthink us, or if "certain elements of human judgment [are] indispensable" in making important decisions (Pazzanese). Moral decisions are made by everyone daily. For instance, when a driver chooses to slam on the brakes to avoid hitting a jaywalker, they are making the moral decision to shift imminent risk from the pedestrian to the passengers in the car. However, delegating moral decisions to AI - something not human - can understandably instill a sense of unease in people. For example, imagine an autonomous car with broken brakes going at full speed towards a grandmother and a child. By deviating a little, one can be saved (UNESCO). The question is whether we can trust automated machines to make those types of decisions for us.

AI has progressed to match the human brain in many areas with increasing accuracy, speed, and quality. Nevertheless, AI algorithms often still "miss the big picture and most times can't analyze the decision with reasoning behind it", according to Joe McKendrick and Andy Thurai, writing for the Harvard Business Review. They shared an incident in which a self-learning bot on Twitter was designed to learn and gather data from interactions with real humans, but instead "it learned offensive language and incorrect facts from other users and didn't engage in proper fact-checking" (McKendrick and Thurai). Microsoft had to take down the bot within 24 hours, acknowledging that it was a "learning experience" with AI. This experience illustrates how AI may not be ready to take on "human qualities that emphasize empathy, ethics, and morality" (McKendrick and Thurai).

In conclusion, the integration of AI into human activities presents significant ethical challenges that cannot be overlooked. As demonstrated, AI systems perpetuate societal biases, compromise individual privacy, and undermine human judgement in moral decision-making. Through analysis of the existing literature, it becomes clear that these issues are not merely technical but are deeply rooted in societal values. Continued research in this area is significant as the development of AI should prioritize fairness, privacy, and human dignity. Addressing such challenges is critical to ensuring that AI serves as an equitable and safe tool for humanity.

Works Cited

Cataleta, Maria Stefania. Humane Artificial Intelligence: The Fragility of Human Rights Facing AI. East-West Center, 1 Jan. 2020. JSTOR, www.jstor.org/stable/resrep25514?searchText=&searchUri=%2Faction%2FdoBasicSearch%3FQuery%3Dai%2Bethics%2Bin%2Bwork%26so%3Drel&ab_segments=0%2Fbasic_search_gsv2%2Fcontrol&searchKey=&refreqid=fastly-default%3A678f436124e7b7fadd04acffb4489f11&initiator=recommender. Accessed 11 Dec. 2024.

Chin-Rothmann, Caitlin. "Protecting Data Privacy as a Baseline for Responsible AI." Center for Strategic and International Studies, 18 July 2024, www.csis.org/analysis/protecting-data-privacy-baseline-responsible-ai. Accessed 15 Jan. 2025.

Gibson, Kate. "5 Ethical Considerations of AI in Business." Harvard Business School Online, 14 Aug. 2024, online.hbs.edu/blog/post/ethical-considerations-of-ai. Accessed 11 Dec. 2024.

McKendrick, Joe, and Andy Thurai. "AI Isn't Ready to Make Unsupervised Decisions." Harvard Business Review, 15 Sept. 2022, hbr.org/2022/09/ai-isnt-ready-to-make-unsupervised-decisions. Accessed 2 Jan. 2025.

Pazzanese, Christina. "Great Promise but Potential for Peril." The Harvard Gazette, 26 Oct. 2020, news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/. Accessed 17 Dec. 2024.

Prunkl, Christina. "Human Autonomy at Risk? An Analysis of the Challenges from AI." Minds and Machines, 24 June 2024, link.springer.com/article/10.1007/s11023-024-09665-1. Accessed 19 Dec. 2024.

UNESCO. 21 Apr. 2023, www.unesco.org/en/artificial-intelligence/recommendation-ethics/cases. Accessed 16 Dec. 2024.

UNESCO. www.unesco.org/en/artificial-intelligence/recommendation-ethics. Accessed 17 Dec. 2024.

Aerodynamic Performance of the Front Wing in Relation to Ground Clearance Values
By Soki Ito

Abstract

The front wing design of Formula vehicles has many aspects of aerodynamic performance. An effective front wing design minimizes drag force and maximizes the downforce, which stabilizes the vehicle and increases its speed. This research paper examines more specifically the effects of ground clearance values of a front wing on the aerodynamics of a Formula vehicle. Computer Aided Engineering (CAE) software is utilized to simulate the condition of the front wing while moving under air. Different ground clearance values are applied to investigate the impact they have on down and drag force, hence the aerodynamics of the car. The results concluded that a low ground clearance value was necessary for the aim to maximize aerodynamic performance.

Introduction

The aim for most race cars is to achieve the highest possible speed within their design. Aerodynamics is a crucial factor that enables the car to maintain a fast speed. For the aim to ensure good aerodynamics, there is a need to minimize drag force and maximize the downforce (Castro and Rana). Drag force refers to the resistive force the vehicle faces, which usually acts against the vehicle's motion in fluids. Drag force is considered unfavorable for the formula vehicle, as it reduces the speed of the car. Drag force can be calculated by the following mathematical equation:

$$F_D = 0.5\rho V^2 C_D A, \quad (1)$$

where F_D is the drag force, ρ is the density of the fluid, V is the speed of the object, A is the cross-sectional area, and C_D is the drag coefficient (Hetawal et al.)

Downforce, on the other hand, is a force that pushes the vehicle to the ground. Downforce is essential for race cars, as it increases the tire's friction, allowing for better acceleration and maneuverability of the vehicle (Roberts et al.). Downforce can be calculated by the following equation:

$$F_L = 0.5\rho V^2 C_L A, \quad (2)$$

where F_L is the Downforce, ρ is the density of the fluid, V is the speed of the object, A is the cross-sectional area, and C_L is the drag coefficient. (Patil et al.)

Ground clearance value (H) is the distance between the ground and the bottom of the front wing. Due to the ground effect, a lower ground clearance will generate a stronger downforce, increasing the stability of the car. (Katz) A lower ground clearance value will also

minimize impact drag force, as it reduces impacts due to unnecessary airflow from underneath the front wing. However, a ground clearance value that is overtly low also raises problems, such as potential safety concerns resulting from the low wing placement. Ground clearance values are often normalized by the chord length (C) of the front wing. This allows for more meaningful comparisons when evaluating the airflow in context with the geometry of the front wing. Thus, many optimizations are performed using *H/C* values. Considering the impacts of ground clearance, we hypothesize that a fairly low *H/C* value is required for the front wing to maximize aerodynamic performance and will investigate that further in this research paper.

Similar optimizations have been applied to the front wing in the past. One common optimization parameter is the angle of attack. The angle of attack is defined as the angle between the plane of the vehicle and the direction of the airflow. Varying the angle of attack affects the vehicle's downforce as well as the drag force, thereby influencing the aerodynamics.

(Wordley and Saunders) have investigated the effect of the angle of attack of the front wing on the aerodynamic performance of a Formula SAE car. They utilized a 2-dimensional CFD analysis in their work, to simulate the relationship between the angle of attack and the force coefficients of the front wing. The simulation was performed in a free-stream airflow condition. The results of his simulation concluded that an angle of attack between 15-20 degrees saw an increasing trend in downforce coefficients. However, utilizing an angle of attack larger than 22 degrees began to face decreasing trends in the downforce coefficient. Conversely, a higher angle of attack led to observations of a significant increase in drag force.

(Szudarek and Piechna) have also considered the importance of the angle of attack on the aerodynamics of a sports car. In their works, CFD simulations were also applied to investigate the impacts of varying angles of attack, considering the effects on downforce and drag force. Within the results of the simulation, Szudarek and Piechna observed that increasing the front wing angle of attack has resulted in strong increases in aerodynamic drag force. Their works also showed how the angle of attack and downforce coefficients had an inverse relationship. The results from their simulation suggested how aerodynamic performance was optimized with a maximized lift-to-drag force ratio when the angle of attack was set to 5 degrees.

Closer to the current study, other studies have also optimized ground clearance. For instance, the impacts of ground clearance have been studied in Kachare's works (Kachare). In his study, Kachare applied CFD simulations for a multi-element front wing, referencing 2015 model Formula One cars to investigate how ride heights of the front wing impact aerodynamics. The results from the simulation suggested that the downforce coefficient increased with lower levels of ground clearance values. This was explained in Kachare's works that a lower ground height creates stronger suction between the car and the ground. The suction enables a faster airflow underneath the front wing, allowing for greater velocity of the car. Downforce was maximized at a ground clearance value of 0.09. His works also evidenced that drag coefficients generally increased with decreasing *H/C* values, however, the magnitude of the increase was rather minor.

(Zerihan and Zhang) in their study have also considered the impacts of ground clearance. In their study, an experiment was conducted to test the downforce and drag force of

single-element rectangular element wings under different ride heights and incident angles. The experiment was run in a 2.1m with 1.7m wind tunnel, with the dimensions of the front wing having a span of 1100mm and a chord length of 223.4mm. Within the results of the experiment, they saw higher downforce coefficients at lower ride heights. However, at an *H/C* value of less than 0.15, the downforce coefficient began to first gradually and then significantly decrease. It was observed that the lift coefficient was maximized at $C_L = 1.72$ at a ride height of 0.08c. On the other hand for drag force, decreasing ride heights led to an increase in drag. Zerihan and Zhang discussed how increased proximity to the ground led to greater induced drag, therefore resulting in greater C_D values.

This paper is structured in 4 sections. In the methods section, we describe how computer-aided engineering software (CAE), SimScale, was utilized to create airflow simulations with varying ground clearance of the front wing. In the results, we present the drag force and downforce that were collected in the simulation with different ride heights of the front wing simulated. In the discussion, we compare and analyze the results obtained to see the trends within ride heights of the front wing and the ratios of the force coefficients (C_L/C_D). The discussion also includes limitations of the simulation concerning potential barriers in real-world applications. Finally, in the conclusion we summarize the research paper and state the ride height that was sufficient to maximize aerodynamic performance.

Methods

To test how ground clearance values will affect drag/downforce, hence the aerodynamic performance of the car, a computer-aided engineering (CAE) simulation, using the software platform Simscale, was applied. The simulation utilized a K-omega turbulence model, which is used to approximate complex Navier-Stokes equations (Bang et al.). This will predict the airflow of the Formula vehicle which helps quantify the down and drag forces. The parameters we are varying are the ground clearance values, which alter the height of the front wing of the formula car. For the simulation, we test the ground clearance normalized by the chord length (*C*), hence *H/C* are the values we change. The parameters we are testing are the downforce and drag force, which allow the evaluation of aerodynamic performance. Such simulations to test the aerodynamic performance due to varying ground clearance heights were constructed through the following procedures:

1) Front Wing Design

The simulations were carried on and variables were calculated using a consistent front-wing design. Figure 1 is the half wing utilized for our simulation, with the dimensions presented in Table 1. The use of a half wing will reduce the computational workload in the simulation, enabling faster processing of results.

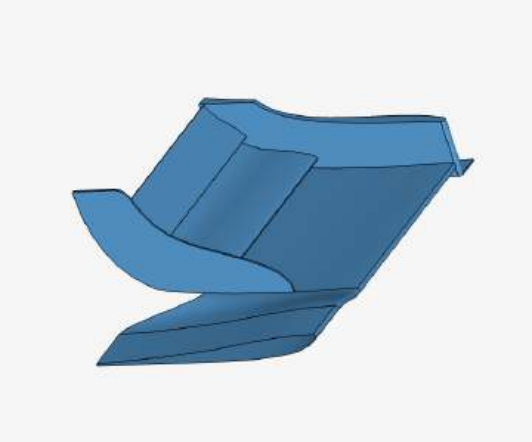


Figure 1: Geometry of front (half) wing

Table 1: Dimensions of front wing

Width (m)	1.46
Length (m)	0.548

2) Assign of flow region

Flow region is essentially a given volume space that records how fluids (air) behave in a particular environment. Flow regions were assigned in the simulation to assess the aerodynamic performance of the front wing. This allows the collection of airflow data and the application of the turbulence model to be assigned within a specific region. By creating a flow region that encloses the front wing, we could test the airflow of the front wing with increased precision and accuracy. Figure 2 is the visualization of the flow region assigned, with the measurement of dimensions as presented in Table 2. ($H/C = 0.085$)

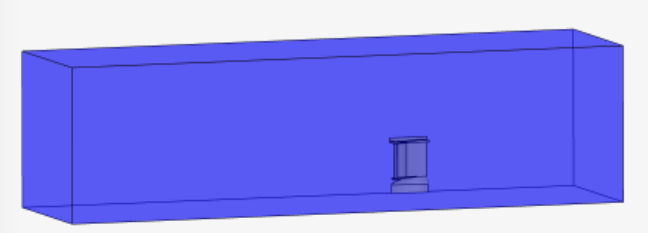


Figure 2: Flow region dimensions

Table 2: Dimension measurement of flow region when H/C = 0.085

Box dimensions (m)			
X min (m)	-6	X max (m)	2.2
Y min (m)	0	Y max (m)	2.2
Z min (m)	0	Z max (m)	2.2

In order to change the ground clearance values, the coordinates for the flow regions were changed accordingly, increasing the ground height of the front wing. This simulation sets the Z-axis as the vertical, which is the height for this context. Adjusting the value for the Z-min value of the flow region changes the elevation of the ground, altering the *H/C* value. For instance, a lower Z-min value would decrease the elevation of the ground, which increases the height of the front wing. This will allow more airflow between the wing and the ground. Likewise, a greater Z-min value would increase the elevation of the ground, which decreases the distance between the ground and the front wing, allowing less airflow.

3) Assign of the conditions of the car

For this research paper is to investigate the impact of ground clearance values, other variables associated with the simulation were kept constant throughout. Some of these variables include the density of the flow material (air for this simulation), turbulence KE, etc. Table 2 summarizes the values set to the variables for the initial condition.

Table 3: Simulation parameters

Air density (kg/m^3)	1.196
Turbulence KE (J)	0.00375
Dissipation rate (1/s)	3.375

Boundary conditions were also assigned and kept constant throughout the simulation. Boundary conditions refer to the conditions set to the walls of the flow region. The main boundary conditions include the velocity inlet, symmetry, pressure outlet, and slip wall. Velocity inlet controls the velocity of the fluid (air for the context of our simulation) that enters the flow region. Similarly, the pressure outlet measures the fluid's flow pressure in the outlet of the flow region. Symmetry allows the flow pattern of the fluid to be mirrored along the surface, which accounts for the simulation with a half wing. Lastly, the slip wall allows the movement of the fluid to remain in a free-stream condition, meaning there is no interaction of the fluid with the boundary walls. Figure 3 illustrates the 3D model for the boundary conditions and Table 4 summarizes the numerical values assigned for the velocity inlet and pressure outlet.

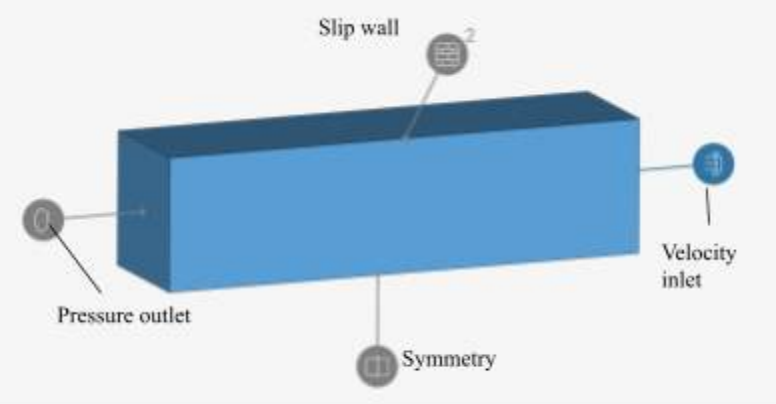


Figure 3: Boundary conditions of the car

Table 4: Boundary conditions for race car

Velocity inlet (m/s)	-15
Pressure Outlet (Pa)	0

4) Refinements for mesh

Mesh refinements were incorporated in the simulation. Meshing simplifies the solving of Navier-stokes equations, as it organizes complex 3-dimensional figures in a computable manner. Increased accuracy in results could be obtained through appropriate mesh refinements.

Cartesian boxes were the main mesh refinement applied to the simulation. Cartesian boxes are essentially assigned regions where a refinement of the mesh is applied to extensively test and approximate the Navier-Stokes equations and provide more accurate results. This process allows further focus on the aerodynamic performance of the front wing. Two layers of cartesian boxes were used (both dimensions smaller than the flow region) to generate more precise results around the front wing region. Similar to the flow region, the Z-coordinates are also adjusted for the cartesian boxes upon varying *H/C* values for consistency and accuracy. The dimensions for the cartesian boxes are presented in Table 5.

Table 5: Dimensions for bigger cartesian box a) and smaller cartesian box b)

a)

Minimum x (m)	-2.5
Minimum y (m)	0

Minimum z (m)	-1.6e-4
Maximum x (m)	0.5
Maximum y (m)	1
Maximum z (m)	0.5

b)

Minimum x (m)	-1
Minimum y (m)	0
Minimum z (m)	-1.6e-4
Maximum x (m)	0.16
Maximum y (m)	0.85
Maximum z (m)	0.25

5) Mesh run

Following the setting of the refinements, the mesh was then run into account. The mesh was completed with a mesh fineness value of 3 (Simscale's mesh fine-coarse scale). Figure 4 demonstrates the 3D model of the simulation after the mesh was completed.

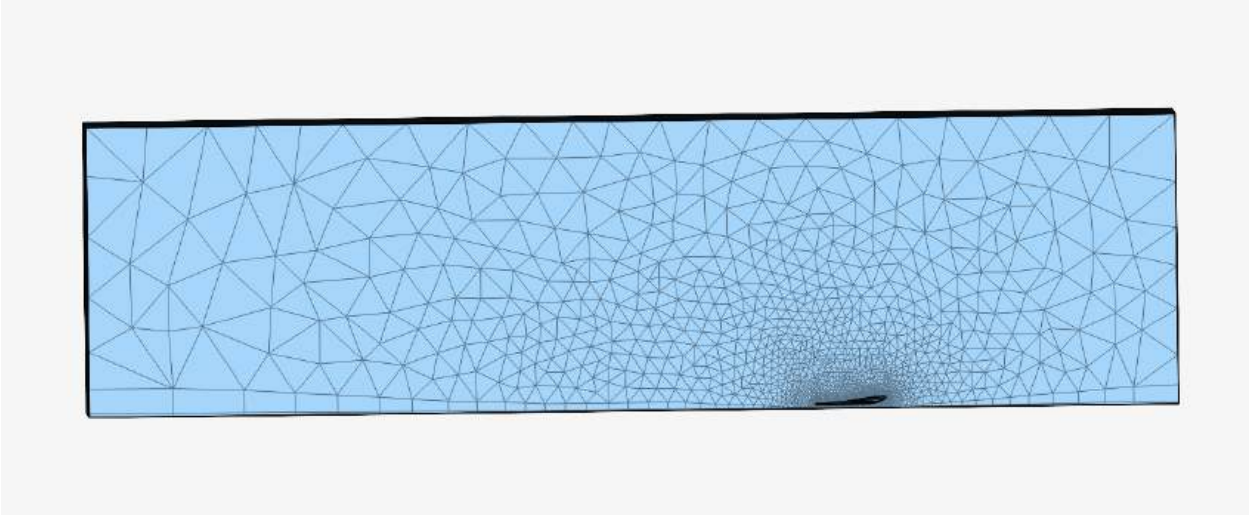


Figure 4: Mesh model (mesh fineness 3)

After the completion of the mesh, the simulation was then run through the K-omega SST turbulence model. The same procedure was conducted for the simulations operating with all ground clearance values, which are $H/C = \ 0.025,\ 0.05,\ 0.085,\ 0.1,\ 0.12,\ 0.15$, and 0.2.

Results

Figure 5 presents the results between *H/C* values and their relative drag and downforces.

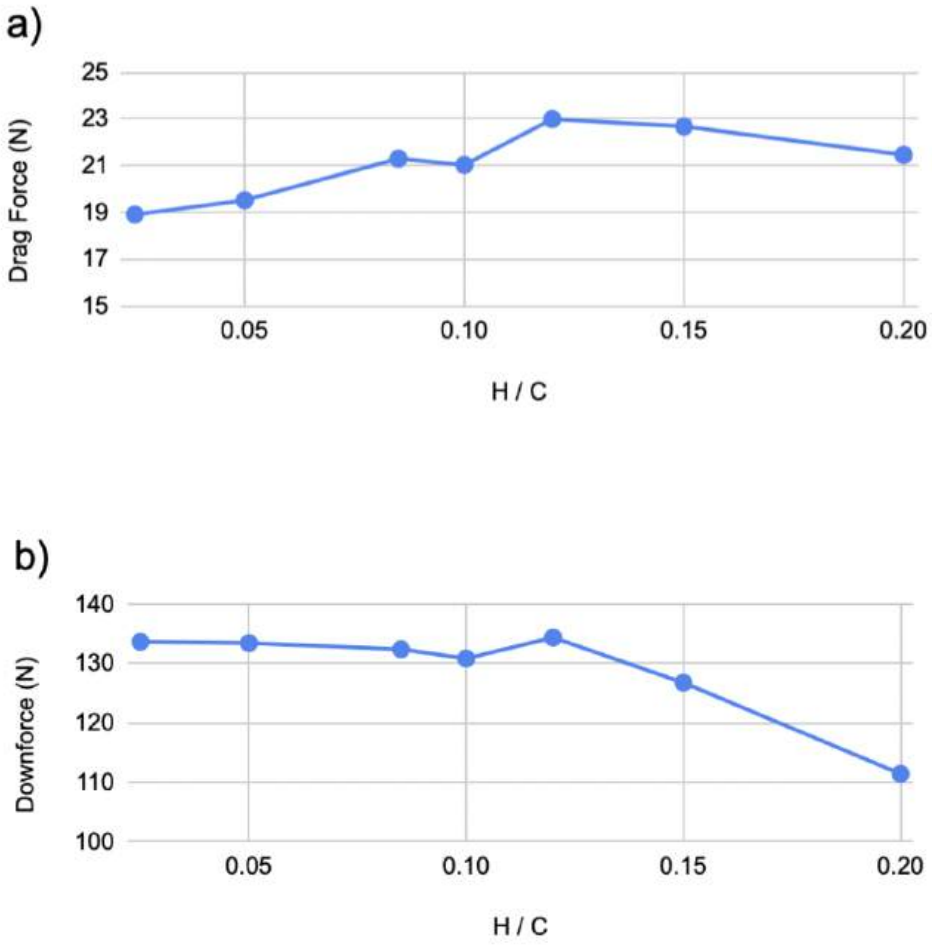


Figure 5: Graphical relationship between H/C values and Drag force (a) and Downforce (b)

Analyzing the results in Figure 5a, there is a general increasing trend in drag force as the H/C values rise. The drag force hits a peak value at $H/C \ = \ 0.12$, where $F_D = \ 22.99\ (N)$. As ground clearance values become greater, more airflow is allowed by the raised front wing height, resulting in the increase of drag force and thus the reduction of aerodynamic efficiency. The trend of increased drag force along with increased *H/C* values seems to support the initial hypothesis. On the contrary, referring to Figure 5b, the downforce of the front wing shows a

general decreasing trend. Downforce remains relatively consistent at lower *H/C* values, however, faces a significant drop from $H/C = 0.12$. The downforce reaches a minimum value of $F_L = 111.41\ (N)$ at $H/C = 0.2$, which is the greatest *H/C* value tested in the simulation. As the elevation of the front wing is raised, there is less ability of the race car to push itself down to the ground, which reflects the decreasing trend of the downforce values.

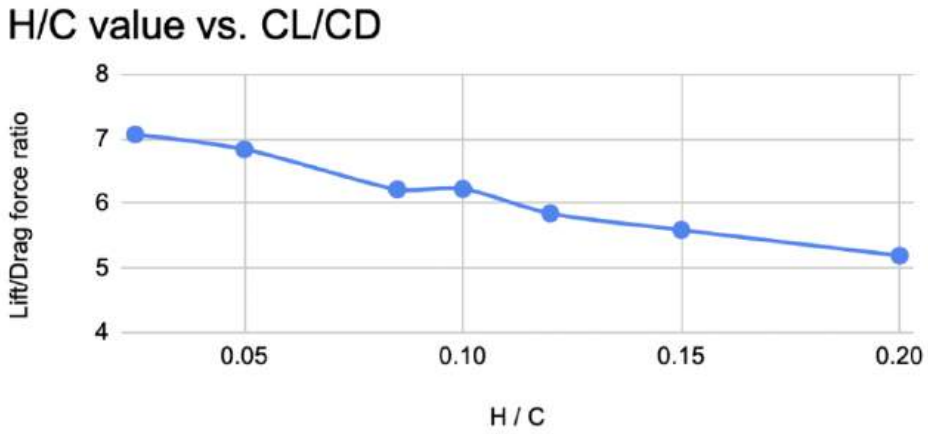


Figure 6: Graphical relationship between H/C values and CL/CD

Seen through Figure. 6, *H/C* value and C_L/C_D ratios show an inverse relationship. As the ground clearance height is raised, the Lift/Drag force ratio is decreased. C_L/C_D reaches its maximum at $H/C = 0.025$ with a value of 7.07.

Discussion

The results of the simulation seem to support and agree with existing literature, and hence the hypothesis that assumed a fairly low *H/C* value was necessary for such intended purpose. For a formula vehicle to achieve good aerodynamic performance, there were 2 key components necessary, which were a maximized downforce and minimized drag force. A greater downforce allows improved maneuverability and stability of the vehicle, and a minimized drag force enables minimal reduction in speed. The combination of the two components will enable a greater C_L/C_D value as presented in Figure 6. Referring to Figure 6, we can conclude that a low ground clearance value of 0.025 is required to achieve optimal aerodynamic performance. With the front wing adjusted to this *H/C* value, the Lift/Drag force ratio will be 7.07, clearly indicating aerodynamic efficiency. The results help convince and validate the hypothesis.

However, such values could not be feasible in practice due to several limitations of the simulation. All simulations do not fully reflect real-life conditions, hence application to actual race cars may not be entirely possible. The simulation we conducted did not consider a full vehicle analysis, hence airflow may change, which impacts the results. Moreover, considering

the elasticity component, changes in the geometry of the front wing may have been neglected. For instance, deformation of the front wing may change the angle of attack, which may potentially influence the C_L/C_D results. Additionally, race cars have front wing constraints set as competition rules. The International Automobile Federation has restricted the height of the front wing in 2022 (Engineering). The FIA suggests that a front wing that is excessively close to the ground can result in safety concerns, due to sudden exceptional drops in downforce. This distorts the race car's aerodynamic stability and significant downforce reduction. Such constraints have been neglected in the case of our simulation.

Conclusion

The original question for this research paper was to analyze the impact of ground clearance values of a front wing and how they impact aerodynamic performance. To investigate the research question, CAE software enabled the simulation of airflow around the front wing with different ride heights. A comparison of downforce and drag force showed that $H/C = 0.025$ will allow maximized aerodynamic efficiency, which generally agreed with the hypothesis. For future work, multimodal simulations could be applied to test further results, such as the consideration of deformation due to component elasticity. Overall, the research paper offered a meaningful comparison of ground clearance values with aerodynamic performance and consideration to real-world applications.

Acknowledgement

I would like to thank Dr. Plinio Zanini for mentorship and consistent supervision throughout constructing this research paper.

Works Cited

Bang, Chris Sungkyun, et al. “Numerical Investigation and Fluid-Structure Interaction (FSI) Analysis on a Double-Element Simplified Formula One (F1) Composite Wing in the Presence of Ground Effect.” Fluids, vol. 7, no. 2, 2, Feb. 2022, p. 85. www.mdpi.com, https://doi.org/10.3390/fluids7020085.

Castro, Xabier, and Zeeshan A. Rana. “Aerodynamic and Structural Design of a 2022 Formula One Front Wing Assembly.” Fluids, vol. 5, no. 4, 4, Dec. 2020, p. 237. www.mdpi.com, https://doi.org/10.3390/fluids5040237.

Engineering, Racecar. “Tech Explained: 2022 F1 Technical Regulations.” Racecar Engineering, 15 Feb. 2022, https://www.racecar-engineering.com/articles/tech-explained-2022-f1-technical-regulations/.

Hetawal, Sneh, et al. “Aerodynamic Study of Formula SAE Car.” Procedia Engineering, vol. 97, Jan. 2014, pp. 1198–207. ScienceDirect, https://doi.org/10.1016/j.proeng.2014.12.398.

Kachare, Shardul C. A CFD Study of a Multi-Element Front Wing for a Formula One Racing Car.

Katz, Joseph. “Aerodynamics Of Race Cars.” Annual Review of Fluid Mechanics, vol. 38, no. 1, Jan. 2006, pp. 27–63. DOI.org (Crossref), https://doi.org/10.1146/annurev.fluid.38.050304.092016.

Patil, Aniruddha, et al. Study Of Front Wing Of Formula One Car Using Computational Fluid Dynamics. 2014.

Roberts, Luke S., et al. Aerodynamic Characteristics of a Wing & Flap Configuration in Ground Effect & Yaw.

Szudarek, Maciej, and Janusz Piechna. “CFD Analysis of the Influence of the Front Wing Setup on a Time Attack Sports Car’s Aerodynamics.” Energies, vol. 14, no. 23, 23, Jan. 2021, p. 7907. www.mdpi.com, https://doi.org/10.3390/en14237907.

Wordley, Scott, and Jeff Saunders. “Aerodynamics for Formula SAE: A Numerical, Wind Tunnel and On-Track Study.” SAE Technical Papers, Apr. 2006. ResearchGate, https://doi.org/10.4271/2006-01-0808.

Zerihan, Jonathan, and Xin Zhang. Aerodynamics of a Single Element Wing in Ground Effect.

The Relationship Between Democracy and GDP By Alina Zhu

Abstract

This paper investigates the complex relationship between democracy, GDP per capita, and GDP growth, providing a nuanced understanding of how governance influences economic trajectories. Using a dataset spanning 151 countries from 1984 to 2021, the study employs regression analyses to explore the economic outcomes. Findings reveal a moderate positive correlation between democracy and GDP per capita, suggesting that democratic systems foster conditions conducive to long-term economic prosperity. However, the relationship between democracy and GDP growth is weaker and more complex, often reflecting inefficiencies such as a slower decision-making and redistribution-focused policies inherent in democratic systems. Regression analyses show that democracy's direct impact of GDP growth is minimal, with broader structural and contextual factors playing a more significant role. These results underscore the importance of considering institutional quality, regional dynamics, and economic policies in analysing the democracy-growth relationship. By highlighting these complexities, the paper contributes to ongoing debates about the role of governance in fostering sustainable economic development.

Keywords

democracy, growth, GDP, GDP per capita

1. Introduction

The intricate relationship between democracy and economic development has long intrigued economists, political scientists and policymakers alike. This paper explores the dynamic interplay between democracy, GDP per capita, and GDP growth, aiming to uncover the nuanced effects of governance systems on economic outcomes. While democratic regimes are often associated with transparency, institutional stability, and the promotion of individual freedoms, their economic impacts remain a subject of debate. By analysing a comprehensive dataset of over 5,000 observations, this study examines the correlations and causal inferences between democratic governance and key economic indicators. What sets my study apart is its use of the most recent data, which allows it to highlight findings that may differ slightly from those of previous research, shedding new light on this ongoing discussion.

The relationship between democracy and economic growth has long been a subject of scholarly debate. While earlier studies highlighted democracy's potential to foster growth and human development, more recent research has revealed a nuanced and, at times, contradictory picture. This review synthesizes findings from three key studies that explore the complex interaction between democracy, economic performance, and crisis management, particularly in the context of the COVID-19 pandemic.

The relationship between democracy and economic growth is a complex and context-dependent one, as highlighted by numerous studies. Heo and Tan (2001), for instance,

used a Granger causality analysis to examine the bidirectional relationship between democracy and growth in 32 developing countries from 1950 to 1982. Their findings revealed that in 34% of cases, economic growth led to democratization, while in 31% of cases, democracy enhanced economic growth, and in 25%, no significant relationship was observed. This underscores the nuanced nature of the democracy-growth dynamic, suggesting that neither variable consistently precedes the other.

Similarly, Doucouliagos and Ulubaşoğlu (2008), in a meta-analysis of 84 studies synthesizing 483 estimates, found no strong direct relationship between democracy and growth. However, they highlighted the importance of democracy's indirect effects, such as human capital accumulation, political stability, and economic freedom. These findings reflect the intricate and multifaceted nature of the democracy-growth link, suggesting that the effects of democracy are more subtle than previously thought. Pourgerami (1988) also emphasized the context-specific nature of this relationship, noting that countries transitioning between autocracy and democracy experience varying economic outcomes depending on institutional stability and political conditions.

Barro's (1996) extensive analysis of data from roughly 100 countries between 1960 and 1990 supported the idea that democratic institutions—such as the rule of law, free markets, and high levels of human capital—positively influence economic growth. However, his study revealed a non-linear relationship between democracy and growth: at low levels of political freedom, democratization spurred growth, but at higher levels, it often hindered it. This was primarily due to democracy's tendency to introduce redistributive policies that prioritize equity overgrowth, such as wealth transfers and the influence of special interest groups.

Building on this, Nagasaki and Nagasaki (2024) argue that democracy's negative effect on growth has become more pronounced in the 21st century. Their instrumental variable analysis of data from 2001 to 2019 found that democracies, particularly in the West, experienced slower GDP growth compared to authoritarian regimes like China. This slowdown, they suggest, is due to factors such as reduced capital formation, slower growth in value-added sectors, and less dynamic international trade. While democracies offer long-term political stability and protect individual freedoms, the authors argue that their political systems may not be as conducive to rapid economic expansion as authoritarian regimes, which can implement economic policies more swiftly and with fewer constraints.

Overall, these studies suggest a complex correlation between democracy and economic growth. While democracy provides essential governance benefits, such as political stability and the protection of individual freedoms, it may also slow down economic growth due to the political compromises and redistributive policies inherent in democratic systems. This is particularly evident in mature democracies, where political freedoms are well-established and the focus shifts toward welfare-oriented policies rather than growth-oriented reforms.

The direct impact of democracy on economic growth remains one of the most debated aspects in the literature, with studies producing contradictory findings. Heo and Tan (2001) show that while economic growth can lead to democratization, and democracy can enhance growth in

certain cases, the relationship is not uniform across all countries. Their results demonstrate a mixed set of outcomes—economic growth promotes democracy in some contexts, and democracy facilitates growth in others, but there are also instances where no significant relationship exists.

Doucouliagos and Ulubaşoğlu (2008) come to a similar conclusion in their meta-analysis, finding that democracy does not have a strong direct effect on growth. Out of the 483 estimates they analysed, the majority showed either no significant correlation or inconsistent findings on the direct impact of democracy on growth. However, their analysis reveals that once indirect mechanisms are considered, democracy plays a more constructive role in fostering growth. Pourgerami's (1988) analysis supports the view that the direct link between democracy and growth is ambiguous. He notes that in certain contexts, democracy can lead to welfare spending, wage rigidity, and pressure from organized interest groups, all of which may slow economic growth.

These inconsistencies are explained by the varying conditions under which democracies operate. For instance, in low-income democracies with weak institutions, the government may be more susceptible to populist pressures, leading to policies that hinder growth, such as excessive welfare spending or inefficient redistributive measures. Conversely, in more advanced democracies with robust institutions, the political stability and rule of law may create a more conducive environment for economic expansion. Thus, the mixed direct effects of democracy reflect the fact that democracies can adopt different economic models depending on their institutional capacities and socio-political circumstances.

Another major finding from these studies is that the effects of democracy on growth are often region- and country-specific. Doucouliagos and Ulubaşoğlu (2008) report that the impact of democracy on growth differs by region, with Latin American democracies experiencing more significant positive effects compared to countries in Asia, where authoritarian regimes have sometimes achieved faster growth. This variation underscores the importance of historical, cultural, and institutional contexts in shaping the democracy-growth relationship.

Heo and Tan (2001) also highlight country-specific outcomes in their analysis of 32 developing countries. They find that the democracy-growth relationship is highly heterogeneous, with some countries exhibiting a positive causal relationship while others do not. This variability suggests that national factors, such as governance quality, institutional stability, and economic structure, play a critical role in determining how democracy affects growth.

Pourgerami's (1988) analysis of Third World countries further emphasizes the cyclical nature of autocracy and democracy in certain regions, where political instability and regime changes create fluctuating economic outcomes. The instability that often accompanies democratization in these countries complicates the relationship between democracy and growth, making it difficult to establish a consistent pattern. These findings suggest that the institutional environment and regional characteristics heavily mediate the effect of democracy on growth, leading to divergent outcomes across different parts of the world.

The relationship between democracy and economic growth is far from settled, with scholars continuing to debate its intricacies. The literature points to a complex and context-dependent interaction between the two variables, where democracy's indirect effects—such as improving human capital, fostering political stability, and securing property rights—often enhance economic growth. However, the direct impact of democracy on growth remains mixed, with significant variation across countries and regions. The studies reviewed here suggest that while democracy is not inherently detrimental to growth, its effects are shaped by the quality of institutions, regional dynamics, and economic policies. As such, any analysis of the democracy-growth relationship must account for the specific political, economic, and institutional context of the countries in question.

The relationship between democracy and economic growth has been a focal point of scholarly debate for decades. Earlier research often emphasises democracy's potential to foster economic growth and human development through mechanisms like improved governance, inclusive institutions, and greater public accountability. However, recent studies paint a more complex and, at times, contradictory picture. This evolving discussion reflects the dual nature of democracy, while it promotes stability and long-term prosperity, it can also introduce inefficiencies, such as slower decision-making and redistribution-focused policies, that may hinder short-term growth.

Ultimately, this paper aims to provide a nuanced understanding of the interplay between democracy, economic development, and growth. The findings reveal a moderate positive correlation between democracy and GDP per capita, suggesting that democratic systems may foster conditions conducive to long-term economic prosperity. However, the relationship between democracy and GDP growth appears weaker and more complex, underscoring the need to consider broader structural and contextual factors. By delving into these complexities, this study contributes to the ongoing discussion on how governance shapes economic trajectories, particularly in the face of global challenges.

The rest of the paper is organised as follows: In the next section, I present the data and the empirical methods. Then in section three, I discuss my results, exploring the complexities of the democracy-growth relationship. Finally, in the last section, I provide some concluding remarks.

2. Data and Methods

2.1 Data

The primary datasets for this study were sourced from the World Bank's World Development Indicators and The Political Research Services (PRSGroup), providing comprehensive coverage of economic and governance indicators across countries. The dataset spans from 1984 to 2021, capturing the economic and political trends over four decades. The dataset covers 151 countries, representing a diverse mix of low-, middle-, and high-income nations.

GDP growth, the dependent variable, is measured as the annual percentage change in gross domestic product, obtained from the World Bank Data. Democracy is quantified using The PRS Group's data, ranging from 0.00 to 6.00. GDP per capita is reported in constant US dollars, sourced from the World Bank.

Table 1: Descriptive Summary Statistics

Variable	Obs	Mean	Std. Dev.	Min.	Max.	Median
Democracy	5,135	3.819395	1.652016	0.00	6.00	4.00
Growth	5,080	3.258716	6.492995	-64.047	153.493	3.5566
GDP per capita	5,050	13184.36	17712.81	163.57	112417.9	4861.89

Table 1 summarises the democracy and GDP per capita variables. Democracy, with 5,135 observations, has a mean of 3.82, a standard deviation of 1.65, and ranges from 0 (least democratic) to 6 (highly democratic), indicating a broad spectrum of governance systems slightly skewed toward democratic regimes. GDP per capita, based on 5,050 observations, averages $13,184.36 with significant variability (SD = $17,712.81), ranging from $163.57 to $112,417.90, reflecting both low- and high-income countries.

The correlation between democracy and GDP per capita is 0.4212 ($p < 0.0001$), a statistically significant positive relationship. This suggests that more democratic countries tend to have higher GDP per capita, though causation cannot be inferred, as factors like institutional quality and economic policies likely mediate this association. The positive correlation highlights democracy's potential role in promoting economic stability and prosperity. The correlation between democracy and GDP growth is -0.0328, with a p-value of 0.0716, indicating a weak negative relationship that is not statistically significant. This suggests no meaningful linear association between these variables. While democracy correlates positively with GDP per capita, its relationship with GDP growth appears more complex. Democratic systems might face challenges such as slower decision-making processes or redistribution-focused policies that could temper short-term growth. Conversely, authoritarian regimes may implement rapid reforms or growth-oriented policies, leading to faster, albeit potentially less sustainable, growth.

2.2 Methods

This study employs quantitative analysis to investigate the relationship between democracy, GDP growth, and GDP per capita. The primary method of analysis is linear regressions, using GDP growth as the left-hand side (LHS) dependent variable and democracy a as a key right-hand side (RHS) independent variable. The regression models also incorporate GDP per-capita as an additional explanatory variable to explore its role in economic development. The dataset consists of over 5,000 observations, capturing key economic and governance indicators across various countries. GDP growth is measured as the annual percentage change in gross domestic product, serving as the dependent variable. GDP per capita is expressed in US dollars, representing economic development on a per-individual basis.

Democracy is quantified on a scale ranging from 0 (least democratic) to 6 (most democratic), reflecting the spectrum of governance systems. The three different types of regression models using panel data. We use a simple regression where GDP growth is regressed on democracy to assess the direct relationship between governance and economic expansions. We then add GDP per-capita as well as country and year dummies to the benchmark analysis. The models are estimated using robust standard errors to account for heteroscedasticity.

3. Results

Table 2: Regression of Growth on Democracy

Linear regression

Number of obs	= 4,894
F(1, 4892)	= 4.04
Prob > F	= 0.0445
R-squared	= 0.0011
Root MSE	= 6.3121

growth	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
Democracy (b)	-.1261393	.0627494	-2.01	0.044	-.2491563	-.0031224
cons (a)	3.896412	.3028673	12.87	0.000	3.302656	4.490168

The regression analyses reveal a complex relationship between democracy, GDP growth, and GDP per capita. The findings show a moderate positive correlation between democracy and GDP per capita, indicating that democratic countries tend to exhibit higher average income levels. However, the relationship between democracy and GDP growth is weak and statistically insignificant in simple regression models. When additional controls such as GDP per capita, country, and year dummies are introduced, the relationship between democracy and GDP growth becomes positive and statistically significant. This suggests that democracy's growth-promoting effects may depend on broader structural factors. Additionally, GDP per capita exhibits a small but statistically significant negative relationship with GDP growth, consistent with the economic theory of diminishing returns in wealthier nations.

Table 2 presents the results of a simple regression of GDP growth on democracy. The coefficient for democracy is -0.1261 ($p = 0.044$), indicating a weak negative relationship. The R-squared value is 0.0011, suggesting that democracy explains only 0.11% of the variation in GDP growth. These results suggest that while the negative coefficient is statistically significant, its magnitude and explanatory power are minimal. This finding likely reflects inefficiencies in democratic systems, such as slower decision-making or redistribution-focused policies, that may dampen short-term growth.

Table 3: Regression of Growth on Democracy and GDP per capita

Linear Regression			Number of obs =		4,852
			F(2, 4849)	=	8.16
			Prob > F	=	0.0003
			R-squared	=	0.0021
			Root MSE	=	6.3005

growth	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
Democracy	-.0654482	.0710699	-0.92	0.357	-.2047775	.073881
GDPpercap	-.0000127	4.45e-06	-2.87	0.004	-.0000215	-4.02e-06
Const	3.833375	.3083507	12.43	0.000	3.228868	4.437882

When GDP per capita is added as an explanatory variable, as shown in Table 3, the coefficient for democracy becomes -0.0654 (p = 0.357), which is statistically insignificant. The coefficient for GDP per capita is -0.0000127 (p = 0.004), indicating a small but significant negative effect. The R-squared value improves slightly to 0.0021. These findings suggest that the inclusion of GDP per capita reduces the significance of democracy, highlighting its mediating role. The negative coefficient for GDP per capita aligns with the concept of diminishing returns, where wealthier nations experience slower growth.

Table 4: Regression of Growth on Democracy, GDP per capita and Country Dummies

Linear regression			Number of obs =		4,852
			F(136, 4714)	=	.
			Prob > F	=	.
			R-squared	=	0.0645
			Root MSE	=	6.187

growth	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
democracy	.5001307	.1505725	3.32	0.001	.2049382	.7953232
GDPpercap	-3.06e-06	.000018	-0.17	0.865	-.0000384	.0000323
Const	.4610342	1.370119	0.34	0.737	-2.225039	3.147108

Tables 4 and 5 present the results of the regression models that include country-specific and year-specific effects. In these models, the coefficient for democracy becomes positive and statistically significant, with a value of approximately 0.46 (p = 0.001). The coefficient for GDP per capita becomes negligible and statistically insignificant, while the R-squared value improves

substantially, reaching 0.1472 in the most comprehensive model. These results suggest that democracy's positive impact on growth emerges only when broader contextual factors, such as regional and institutional variables, are accounted for. This highlights democracy's indirect contributions to economic performance through mechanisms like political stability and institutional quality.

Table 5: Regression of Growth on Democracy, GDP per capita and Country and Year Dummies

```
Linear regression                               Number of obs =      4,852
                                                F(173, 4677)  =          .
                                                Prob > F      =          .
                                                R-squared     =     0.1472
                                                Root MSE      =     5.9306

------------------------------------------------------------------------------
             |               Robust
      growth | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
   democracy |   .465997   .146672     3.18   0.001     .1784508    .7535432
   gdppercap |   3.03e-06   .0000235    0.13   0.897    -.000043     .000049
       _cons |   3.695234   1.791367     2.06   0.039     .1833104    7.207157
------------------------------------------------------------------------------
```

The correlation between democracy and GDP per capita is 0.4212 ($p < 0.0001$), indicating a moderate and statistically significant positive relationship. This suggests that democratic governance fosters conditions conducive to long-term economic prosperity, such as transparency, investment, and institutional stability. In contrast, the relationship between democracy and GDP growth is weaker, with a correlation of -0.0328 ($p = 0.0716$), indicating no meaningful linear association. The weak relationship between democracy and GDP growth may reflect trade-offs between slower decision-making processes in democratic systems and the rapid policy implementation often seen in authoritarian regimes. Additionally, the lack of a significant correlation highlights that democracy alone does not directly influence the rate of economic expansion. Instead, economic growth likely depends more on structural factors such as industrialization, global trade dynamics, and technological innovation than on governance type.

The regression analysis provides further clarity on these relationships. As shown in Table 2, a simple regression of GDP growth as a function of democracy reveals a small, negative, and statistically significant effect, with a coefficient of -0.1261 ($p = 0.044$). However, the R-squared value of 0.0011 indicates that democracy alone explains only 0.11% of the variation in GDP growth. Adding GDP per capita to the model, as shown in Table 3, diminishes the significance of democracy, with its coefficient dropping to -0.0654 ($p = 0.357$). Meanwhile, GDP per capita shows a slight but significant negative impact on growth, with a coefficient of -0.0000127 ($p = 0.004$). The R-squared value increases marginally to 0.0021, reinforcing the idea that democracy's direct influence on growth is weak and context-dependent. The negative coefficient

for GDP per capita aligns with the concept of diminishing returns, where wealthier countries tend to grow more slowly.

When extended models include numerous control variables, as shown in Tables 4 and 5, the results offer a more nuanced perspective. In these models, the R-squared improves significantly, reaching 0.1472, indicating that the additional factors explain 14.72% of the variation in GDP growth. The coefficient for democracy becomes positive and statistically significant, with a value of approximately 0.46 ($p = 0.001$), suggesting that its role in supporting growth becomes evident when accounting for other variables. Conversely, GDP per capita's coefficient becomes negligible and statistically insignificant, implying that its influence diminishes when broader factors, such as regional or institutional variables, are considered. These findings highlight the complexity of the democracy-growth relationship, suggesting that democracy contributes positively to growth when interacting with other elements of governance and economic structure.

The analysis underscores the complexity of the democracy-growth relationship. Democracy positively correlates with GDP per capita, highlighting its role in fostering economic stability and prosperity. Its direct impact on GDP growth is less pronounced, with broader structural and institutional factors mediating its effects. Wealthier nations experience slower growth, suggesting that economic maturity influences growth trajectories more than governance type. The low explanatory power of the models suggests that other variables, such as trade, technology, or regional dynamics, likely play more significant roles in determining GDP growth. Additionally, the analyses focus on correlations and do not establish causation. Future research could employ instrumental variable approaches to address potential endogeneity issues. The heterogeneity in democracy's effects across regions and income levels warrants further investigation, particularly in developing versus developed contexts. By providing a nuanced understanding of these dynamics, this study contributes to the broader discourse on governance and economic development.

4. Conclusion

This study provides a nuanced exploration of the intricate relationship between democracy, GDP per capita, and GDP growth. While democratic regimes are often linked to transparency, institutional stability, and individual freedoms, their economic impacts are complex and multifaceted. The finding reveals a moderate positive correlation between democracy and GDP per capita, suggesting that democratic systems may foster conditions conducive to long-term economic prosperity. However, the relationship between democracy and GDP growth is weaker and more complex, reflecting the dual nature of democratic governance. By incorporating robust regression analyses and a comprehensive dataset, this paper highlights the need to consider broader structural and contextual factors in understanding how governance systems shape economic trajectories. These insights contribute to the ongoing discussion on democracy's role in fostering sustainable economic development.

Works Cited

Barro, R.J., 1996. Democracy and growth. Journal of economic growth, 1, pp.1-27.

Doucouliagos, H. and Ulubaşoğlu, M.A., 2008. Democracy and economic growth: a meta‐analysis. American journal of political science, 52(1), pp.61-83.

Heo, U. and Tan, A.C., 2001. Democracy and economic growth: A causal analysis. Comparative politics, pp.463-473.

Nagasaki, K., & Nagasaki, S. (2024). Negative impact of democracy on GDP annual growth rate in 2001-2019 and 2020 and mortality rate due to COVID-19 in 2020. International Journal of Applied Economics, Finance and Accounting, 19(1), 133-148.

Pourgerami, A., 1988. The political economy of development: A cross-national causality test of development-democracy-growth hypothesis. Public choice, 58(2), pp.123-141

What Comes Next? A Technical and Fundamental Analysis of the Semiconductor Industry

By Justin Lim and Reid C.

Abstract

The main focus of this paper would be a technical and fundamental analysis of the semiconductor industry to anticipate its potential. Thus, the driving question of the paper is how semiconductors will succeed in the future. The first part of the paper introduces and summarizes the idea of technical and fundamental analysis. For instance, ratios like EPS and P/S are introduced as well as stock patterns and census estimates. Subsequently, the current trend of the semiconductor industry is revealed: the AI and technology market is thriving, different key players are holding their position worldwide, and by applying fundamental analysis. Despite their current trend, the semiconductor industry's challenges are presented. By looking at geopolitical, natural, and technological aspects, the industry is struggling to perceive tension between the US and China, demand more REE, and meet technological equilibrium. Finally, by depicting one of the biggest companies in the semiconductor industry, Nvidia, technical and fundamental analysis is done to reflect on the market's long-term sustainability. Heeding these factors, the future of the world's most flourishing market will be anticipated!

Introduction

The semiconductor industry is experiencing an ongoing global boom, with microchips being one of the most crucial technological components in the modern world. However, following this rise in significance, there are many concerns regarding the industry's long-term stability. Specifically, many assume such a rise in a short period to be a bubble. Nonetheless, the industry has grown due to increased demand due to the rise of artificial intelligence, the development of digital-based automobiles, and the Internet alongside social networks. Simultaneously, stock values and massive emerging manufacturing R&D research and development) budgets strike an all-time high in the market. Held by big players like Nvidia, Intel, Samsung, and NEC, these global companies each play a significant role in specific areas to ensure a harmonized industry. On the other hand, the drawbacks included supply shortage due to lack of REE (rare earth elements), geopolitical tensions between the US and China preventing global semiconductor trades, and technological threats. Due to these challenges that it might encounter, it is difficult to anticipate the prosperity of the semiconductor industry.

Principles of Technical Analysis

Before delving into the details of the semiconductor industry, it is crucial to understand the general market trends of any sector: fundamental and technical analysis. Fundamental analysis considers economic factors that impact industries based on different statistical and numeric metrics.

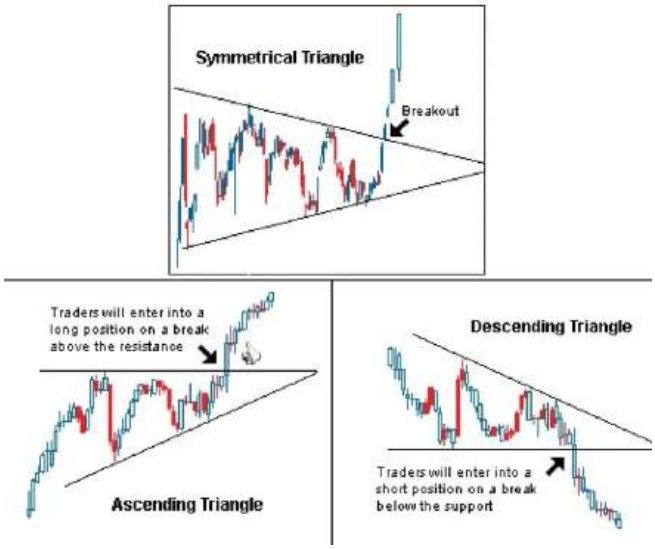
Estimated overall earnings indicate usefulness in the industry's healthy growth. EPS (earnings per share) is the income divided by the number of outstanding shares, denoting the

industry's profitability changes. Free cash flow shows how well the company handles cash. ROE (Return on Equity) depicts how efficient the company is, which is calculated by dividing net income by equity amount. The price-earning ratio (P/E ratio) shows how market price relates to the company's earnings, calculated by dividing earnings by EPS. The price-to-sales ratio (P/S ratio) is calculated by dividing a company's market capitalization by its total revenue.

Second, technical analysis, which analyzes charts and consensus, is an alternative way to examine the trend many analysts use. Since the market never moves in one direction forever, re-established trends or significant changes in the trend either reverse or continue a movement. Some pattern identifications are followed.

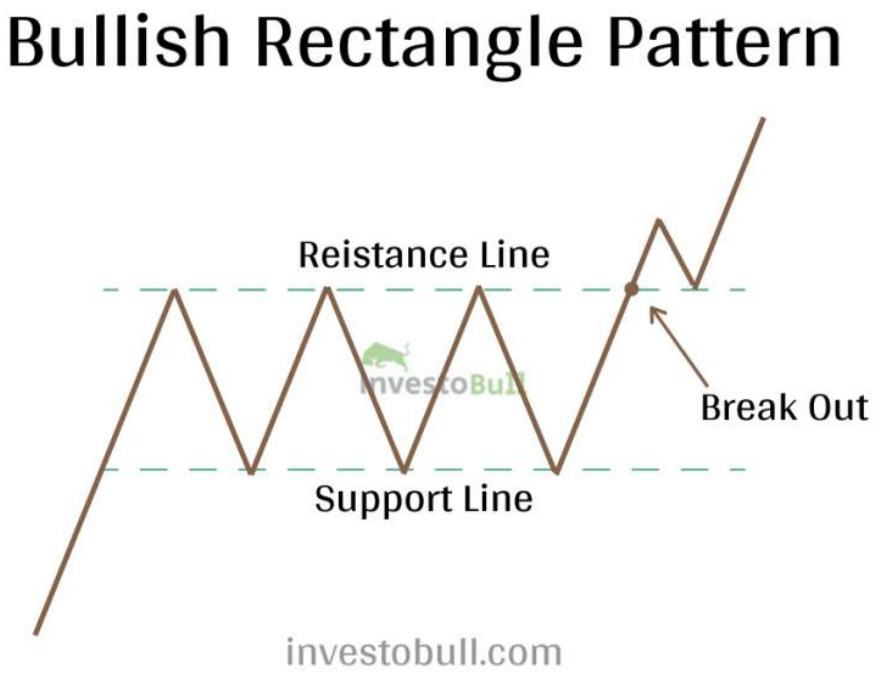
A triangle pattern is identified by drawing two lines from the price bars that look like a triangle. Then, within the triangle, a short range of prices is formed. Subsequently, by looking at the shapes of the triangles, we can indicate the direction of breakouts.

Ascending triangles are bullish, descending triangles are bearish, and symmetrical triangles are directed to either change. For instance, when encountering an ascending triangle when looking at different stocks, it would be beneficial to anticipate growth.



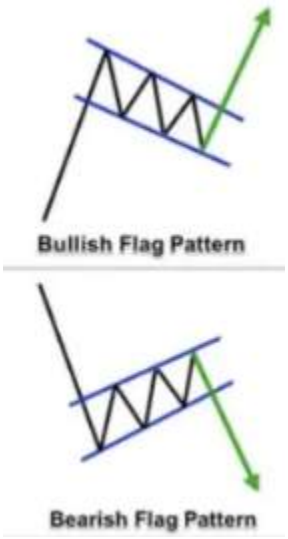
Source: Corporate Finance Institute (28)

A rectangle pattern occurs when the price consolidates. The price moves up and down in a narrow range, forming a rectangle. When the rectangle is formed, it indicates that the price will tend to exit the consolidation in the same direction as it entered.



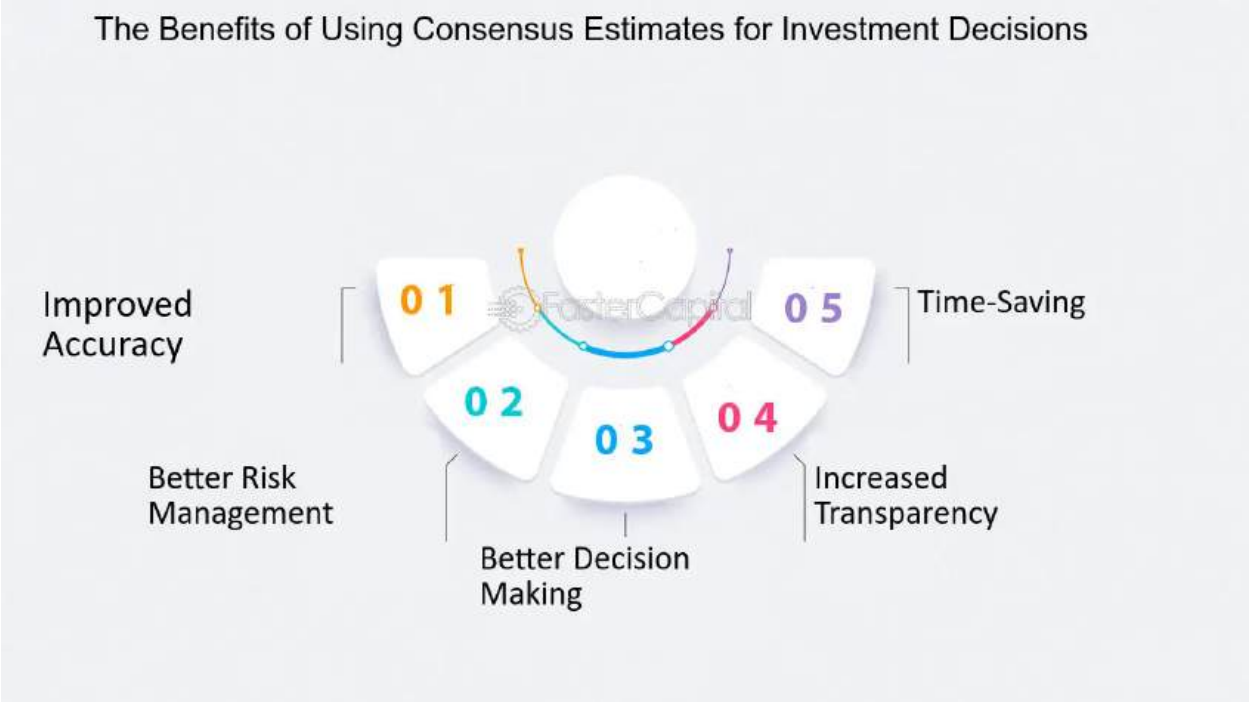
Source: Investobull (29)

A flag pattern is formed when prices are held back in a tight, parallel range for a short period. It also consolidates and resumes the direction that it entered. Flags are seen in mid-trends, not tops or bottoms.



Source: Medium (17)

Subsequently, consensus rating estimates indicate a company's buying and selling rating using various factors and tools. We can use these ratings to anticipate whether the stock is undervalued or overvalued.



Source: Faster (27)

Current state of the semiconductor industry

Not surprisingly, semiconductors aren't the only industry experiencing a boom. Considered as the period of technological development, the twenty-first century was an epoch of many technological booms. These include mobile cellular devices, computers, web browsers, and robotics. For that reason, microchips and semiconductors were a boom within all of these technological devices. Subsequently, this technological boom has been conserved until now, and it has positioned itself towards artificial intelligence, advanced robotics, and automotive. Nonetheless, these developments are leading the boom to the next level, requiring vast amounts of highly efficient chips at the all-time greatest demand rate; just in the year before 2023, there were total global sales of 1 trillion semiconductors. Moreover, the development of AI is exponential, and so is the demand and growth of semiconductors, as "the growth is driven by a continued surge in AI-related semiconductor demand and recovery in electronic production" (1). Specifically, TSMC is set to sustain its strong growth after the report that forecasted a 54% jump in quarterly profit driven by increasing demand for chips used in artificial intelligence. Since TSMC is one of the most prominent leaders in the semiconductor industry, the effects of TSMC's demand for chips would be followed by the growth of many other companies as well (14).

While AI is thriving as the cornerstone of technological future success, the contribution of advanced robotics to semiconductor success cannot be ignored. Due to such precise and efficient labor required in semiconductor manufacturing facilities, many big industry players highly value its precision and efficiency. Regarding this high application, “the Robotics industry is projected to grow at an annual rate of 10.5% and surge its supply chain” (11). Subsequently, the global semiconductor sales revenue has increased due to robotics use, with about 6% growth in revenue in one year (18). Moreover, due to the rise in technology-based automobiles, it is now crucial for cars to have semiconductors. For instance, during 2021 and 2022, the automobile semiconductor sector grew by over 25% and had an 18.1% continuous annual growth rate over time (3). With such constant demand and sector growth, the semiconductor industry will flourish consequently. Yet, a consistently growing sector relates to all these other sectors: the Internet of Things. The Internet of Things, or simply IoT, is a network that connects physical objects and exchanges data between them. As the global IoT market comes close to surpassing 1 trillion dollars by the end of 2024, according to an analysis by GlobalData, there has been significant growth in areas like enterprise IoT, home automation and security, and retail analytics and solutions; regarding the development in these areas, there will be an increase in demand for their micro-controllers, sensors, and memory chips (20).

So, who leads the semiconductor industry? The semiconductor sector is considered one of the most globally sustainable sectors, comprising many major companies globally. Looking into a few of them, TSMC, located in Taiwan, is the biggest manufacturing specialized semiconductor company that yields 62% of the global market share of the foundry (manufacturing) market. TSMC’s dominance in the manufacturing sector benefits the entire market by stabilizing the supply of chips worldwide. Nonetheless, competing in such a district is impossible, obscuring the external companies (23). It's one of the main profit segments from general advanced nodes, responsible for 67% of TSMC’s overall revenue in 2023 (19). Samsung, a South Korean company founded in 1969, is a global leader in technology and electronics, and it has different sectors, such as consumer electronics, information technology, and device solutions. Thereby, Samsung, the second-largest foundry globally, has a market share of 7.5% as of 2023. Its integrated device manufacturing, IDM, allows it to design and manufacture its chips, which also produces semiconductors for other companies such as IBM and Qualcomm (25); Intel, another big player in chip models, has the acquisition of about 9.1% in global market share as of 2023. However, since Intel has failed to follow either of the current trends in the semiconductor industry of developing the fabless model or outsourcing the manufacturing, it is going on a downward path along with the declined sales, experiencing a 30% decrease in revenue this year in 2024; on the other hand, being one of the most successful companies in the world with market capitalization of 3 trillion dollars, Nvidia is a major company focusing in GPU (graphics processor units) component of semiconductors. Specifically, Nvidia has 88% of the total market share in the GPU industry and controls 70% to 90% of the market share for global AI intelligence chips (8). Because various companies worldwide take significant roles, their mutualism is considered a key part of the semiconductor industry.

Applying Fundamental Analysis to Semiconductors

With such boomings and well-coordinated positions of companies, the semiconductor market is also following the economic trend. By applying fundamental analysis of the stock market, we can analyze the current position of the stock market and its profits. The P/E ratio, comparing the market price to the company's earnings, of the U.S. Semiconductor industry is 70, which is almost two times greater than the last three-year average P/E ratio of 38.1 (26). However, a higher P/E ratio isn't always optimal; it is a possible indicator of a stock's overvaluation. This is because the higher the price is relative to the company's earnings, the harder it is for the company to keep up such earning anticipations of such stock price. Nonetheless, a high P/E ratio would indicate great interest and confidence in the industry. On the other hand, the price-to-sales ratio, showing how much value is possessed within every dollar of the company's sales, of the current Semiconductor Manufacturing sector is about 5.47. Although a lower price-to-sales ratio considers the price "undervalued," many analyses still consider it undermined (24). Likewise, within only 5 years, the stock price of Nvidia increased by a shocking 2594.26%, as did TSMC, with an increase of 271.84%.

Challenges

Despite the market's vigorous response to the booming trend of semiconductors, the industry itself is confronting tough challenges. Beginning with the basics, there is currently a supply chain shortage resulting from excess demand and requests for excessive-quality that requires REE (rare earth elements). Especially due to the application of AI in smartphones and PCs, their sales are anticipated to experience 15% and 31% annualized growth, respectively, which requires semiconductor fab to increase their output by 25% to 35%, costing about 40 to 75 billion dollars (12). Furthermore, generating a semiconductor requires many skills, such as precision and brilliance, and foremost, scarce materials, such as rare earth elements (REE). For instance, Germanium and Gallium, due to their unique properties, such as high power density and high switching frequencies, are elements that are critical to many high-tech device manufacturing but are extremely rare on Earth. In addition, the global market for Germanium anticipated its shortfall with a deficit of up to 38 tons in 2023 due to increasing demand from the semiconductor industry (30). Nonetheless, the industry not only has to fight with nature's constraints and development of technology but also geopolitical tensions between the US and China. First, the US and Chinese governments have a trade war and restrictions. Currently, the US controls putting sanctions on advanced semiconductor exports to China, which has led manufacturing companies to reexamine their supply chains and fulfillment strategies (22). Second, the possibility of China invading Taiwan would devastate global supply chains. There are many motives for China to invade Taiwan; for instance, in historical factor, China views Taiwan as a land that has to be reunited with the mainland.

Furthermore, strategic location and crucial role in the semiconductor industry increases China's interest in invasion. As a significant role in the semiconductor industry, Taiwan has led advanced chip technology, with a 68% global capacity share and 80% market share in UV generation

processes (6). A crucial semiconductor company located in Taiwan, TSMC faces many challenges due to the potential invasion threats. Third, massive demand loss was found due to such massive market demand in China. Thus, the World Semiconductor Trade Statistics shows a notable 10% drop in the production of chips and revenue growth due to reduced production levels caused by the tension between the US and China (15).
Third, competing export/import bans are affecting semiconductor production. Many countries, such as the Netherlands, Japan, and South Korea, started following the U.S. in restricting the export of semiconductor manufacturing equipment to China. On December 21, 2023, they also announced the ban on rare earth extraction (10). The impact is significant since China produces up to 60% of all the global REEs (4).

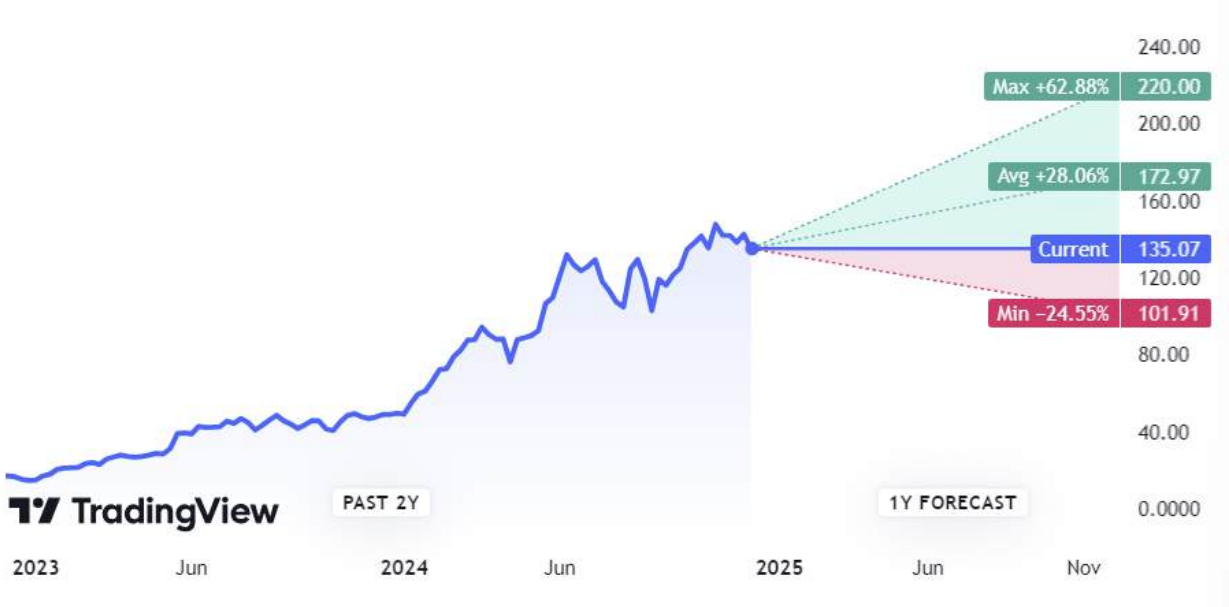
Moreover, the fast-paced development of technology boosts the semiconductor market and simultaneously delivers a threat. First and foremost, finding a "technological equilibrium" is complicated, which means the industry has problems balancing production costs and performance gains. Due to the fast rate of our current innovation, it is almost impossible to generate consistent cost benefits (2). An example can be directed from Moore's law, which states that the number of transistors on an integrated circuit can only double every two years with minimum cost; shown with Moore's law, there is always a limitation to how much technology can develop within a period with minimum cost. Similarly, maintaining the balance between research and development, specialized production, and general processing needs is a great challenge for many companies due to changes in demand. Thereby, it could lead companies to shortages or overproduction of supply. For instance, the industry survey shows that about 51% of the semiconductor companies have delayed capital expenditure due to fast-paced industry and cost pressures. This led to high risk regarding disruption in mass production (7).
Besides, building and maintaining the fab (semiconductor fabrication plant) is followed by extreme cost, requiring up to 10 billion dollars in investment. These abundant costs are for the maintenance of equipment and facilities and the consistent need for highly skilled labor. Due to these high entry barriers, only a few major players in the industry are concentrated on giving supplies, which has risks of causing supply chain disruption.

The long-term strength of the semiconductor

Despite this, the semiconductor industry's inevitable strength in technological dependency will help it succeed in the long term. Last year, in 2023, there was a total sale of 1 trillion semiconductors globally, with 527 billion dollars worth of sales. Regarding the current human population of 8 billion, the number reaches more than 100 semiconductors per person (13). This significant number of sales also represents how much economic impact it will yield. Also, the dominant sales of semiconductors aren't temporary. 6 to 8% of compounded annual growth rate is expected in the industry due to increasing demand in automobile, data storage, and wireless industries (5). One of the most common instances of this would be increasing smartphone use. Globally, there are 5.75 billion unique mobile users, increasing by 2.1 percent yearly.

Moreover, heeding the technological equilibrium problems, semiconductor companies strive for innovation. Many major semiconductor companies invest significantly in their Research and Development sector to acquire high technologies that meet the demanded quality. Consequently, semiconductor companies invest about one-fifth of their revenue annually in R&D.

Furthermore, a technical analysis of recent images of major companies like Nvidia will reveal trends in the semiconductor market.

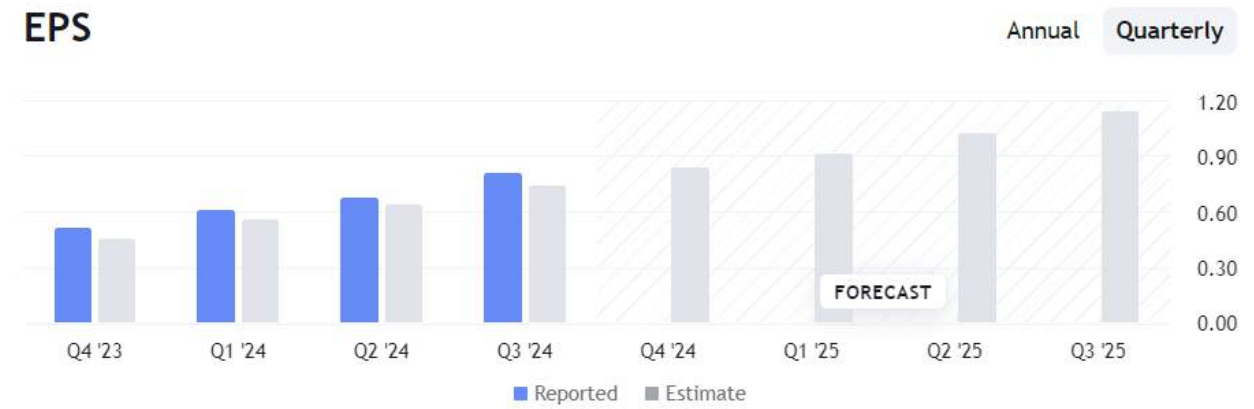


Source: TradingView (16)

The graph on the top depicts Nvidia's current stock price. Looking at the past year, many recent concerns and challenges were presented regarding the semiconductor industry, reaching about an 820% increase in stock price in the last two years. Moreover, inspected by 51 expert analysts in *TradingView,* the maximum and minimum estimated prices of the 1-year forecast are $220 and $101.91, respectively. Heeding that the current price is in about the $130 to $140 range, the average price they suggest is $172.97, which is still about 30% growth within a year. These analysts' forecasts would control the consumer's behavior to purchase the stocks more, which might signal the start of an economic trend.

Source: TradingView (16)

The following image is a consensus of Nvidia stock. It shows a strong buy sign, as indicated by 66 analysts. Such strong buy signals that the stock is in a good state with anticipations to increase its future value.



Currency: USD	Q4 '23	Q1 '24	Q2 '24	Q3 '24	Q4 '24
Reported	0.52	0.61	0.68	0.81	–
Estimate	0.46	0.56	0.65	0.75	0.84
Surprise	+12.34%	+9.17%	+5.39%	+8.26%	–

Source: TradingView (16)

EPS (earnings per share) shows how much earnings are generated for one share. Although the stock is currently highly valued and people anticipate more when comparing quarterly reported and estimated EPS, the reported EPS always exceeds the anticipated EPS. Indeed, a High EPS indicates that the stock is very demanding and profitable.

Currency: USD	Q4 '23	Q1 '24	Q2 '24	Q3 '24	Q4 '24
Reported	22.10 B	26.04 B	30.04 B	35.08 B	–
Estimate	20.40 B	24.62 B	28.78 B	33.17 B	37.96 B
Surprise	+8.37%	+5.78%	+4.38%	+5.76%	–

Source: Faster (16)

Finally, Nvidia's revenue also easily exceeded its anticipated values, generating about two billion dollars more each quarter than the estimate. Revenue is also one of the most crucial factors to consider when discussing the company's state. Without stable and high revenue, the company finds it extremely difficult to maintain its growth. Nonetheless, Nvidia's revenue exceeds expectations and doesn't stop anytime soon.

Public subsidies for AI systems allow the semiconductor industry to thrive as well. For instance, in 2022, the CHIPS and Science Act authorized 52 billion dollars in subsidies for the semiconductor industry. This act allowed the semiconductor industry to exceed all U.S. industries and sustain itself (9). Moreover, the U.S. forecasted a triple of its semiconductor fabrication capacity and is looking forward to growing its size in manufacturing. Specifically, they announced 90 new manufacturing projects, equivalent to a $450 billion investment across 28 states (13).

The most critical reason the semiconductor industry will thrive is that there are no competitors. Even though there have been attempts to develop technological fields, no technology has replaced semiconductors' fundamental role in powering electronic devices (21). Essentially, semiconductors are the cornerstone of technology and the devices most integrated into modern technologies (31).

Conclusion

Overall, the semiconductor industry has numerous challenges and extensive potential. Its excellent growth is fueled by advancements in AI, robotics, automotive, and the Internet of Things, which shows how much the industry contributes to shaping the future generation. Major role players such as TSMC, Samsung, Nvidia, and Intel depict their mutualism and strategic positioning for the industry's global boom.

Nonetheless, these booms are followed by challenges. Supply chain disruption, geopolitical tension, and the rising cost of maintaining fabrication facilities threaten the industry's sustainability. The dependency on REE (rare earth elements) and international restrictions increases the sector's vulnerability. Moreover, the difficulty of locating a balance between rapid technological innovation and efficient development pressures semiconductor companies.

Even so, the long-term outlook of the industry is optimistic. Having over 1 trillion units of sales in 2023 with an expected compound growth rate of 6-8% per year, its technological significance is proven. By having continuous innovation, a crucial investment in the research and development sector, and using technical and fundamental analysis to forecast trends, the semiconductor industry is suited for sustainable and consistent growth in the global market. Due to humans' extreme dependence on technologies, semiconductors invariably would be the key to our digital generation.

Works Cited

"Ai To Drive Global Fab Market Grow Strongly This Year: Report." *The Economic Times*, economictimes.indiatimes.com/industry/cons-products/electronics/ai-to-drive-global-fab-market-grow-strongly-this-year-report/articleshow/114713675.cms. Accessed 28 Dec. 2024.

Aizcorbe , Ana, and Samuel Kortum. "Moore's Law and the Semiconductor Industry." *BEA*, 13 Dec. 2004, www.bea.gov/index.php/system/files/papers/WP2004-10.pdf.

Ashar, Nimish. "Briefcase: Automotive Semiconductor Industry Sees Long-Term Positive Outlook." *S&P Global Mobility* , 5 June 2024, www.spglobal.com/mobility/en/research-analysis/briefcase-automotive-semiconductor-industry-positive-outlook.html.

Baskaran, Gracelin. "What China's Ban on Rare Earths Processing Technology Exports Means." *CSIS*, 8 Jan. 2024, www.csis.org/analysis/what-chinas-ban-rare-earths-processing-technology-exports-means#:~:text=What%20China%27s%20Ban%20on%20Rare%20Earths%20Processing%20Technology%20Exports%20Means,-Photo%3A%20Dilok%2FAdobe&text=China%20announced%20a%20ban%20of,economic%2C%20and%20rare%20earth%20security.

Burkacky, Ondrej, et al. "The Semiconductor Decade: A Trillion-Dollar Industry." *McKinsey & Company*, McKinsey & Company, 1 Apr. 2022, www.mckinsey.com/industries/semiconductors/our-insights/the-semiconductor-decade-a-trillion-dollar-industry.

Chiang, Sheila. "Taiwan Plays a 'Very Crucial Role' in the AI Supply Chain, Says Taiwan Stock Exchange CEO." *CNBC*, CNBC, 19 Apr. 2024, www.cnbc.com/2024/04/19/taiwan-plays-very-crucial-role-in-ai-supply-chain-says-twse-chief.html#:~:text=Taiwan's%20clip%20 dominance,8%25.

Clark, Lincoln, et al. "Global Semiconductor Industry Outlook 2024." *KPMG*, kpmg.com/kpmg-us/content/dam/kpmg/pdf/2024/global-semiconductor-industry-outlook.pdf. Accessed 29 Dec. 2024.

Figliolini, Jason. "Here Is Why Nvidia Chips Are Taking over the Global Semiconductor Market." *ReluTech*, 11 Jan. 2024, relutech.com/blogs/hardware/nvidia-chips-taking-over-global-semiconductor-market-2#_The_Applications_Of.

Forbes, Steve. "How Government Subsidies Are Hurting the Semiconductor Industry." *Forbes*, Forbes Magazine, 29 Feb. 2024, www.forbes.com/sites/steveforbes/2024/02/29/how-government-subsidies-are-hurting-the-semiconductor-industry/#:~:text=In%202022%20 Congress%20 passed%20the,subsidies%20for%20the%20 semiconductor%20industry.\

Freifeld, Karen. "Exclusive: New US Rule on Foreign Chip Equipment Exports to China to Exempt Some Allies | Reuters." *Reuters*, 31 July 2024,

www.reuters.com/technology/new-us-rule-foreign-chip-equipment-exports-china-exempt-some-allies-sources-say-2024-07-31/.

"Global Industrial Robotics Market to 2030: Increasing Trend of Mass Production Drives Growth." GlobeNewswire News Room, Research and Markets, 24 Apr. 2023, www.globenewswire.com/news-release/2023/04/24/2653198/28124/en/Global-Industrial-Robotics-Market-to-2030-Increasing-Trend-of-Mass-Production-Drives-Growth.html.

Kalra, Yamini. "AI Chip Demand: Semiconductor Industry Faces Pandemic Déjà Vu." *CIO*, 14 Nov. 2024, www.cio.inc/ai-chip-demand-semiconductor-industry-faces-pandemic-deja-vu-a-26808#:~:text=A%20 demand%20 increase%20 of%2020,sales%20between%202023%20and%202026.

LaRocca, Greg. "2024 State of the Industry Report Underscores Opportunities and Challenges for U.S. Chip Industry." *Semiconductor Industry Association*, 12 Sept. 2024, www.semiconductors.org/2024-state-of-the-industry-report-underscores-opportunities-and-challenges-for-u-s-chip-industry/.

Lee, Yimou, et al. "TSMC Bullish on Outlook as AI Boom Blows Q3 Profit Past Forecasts | Reuters." *Reuters*, 17 Oct. 2024, www.reuters.com/technology/tsmc-set-report-strong-profit-driven-by-ai-boom-2024-10-16/.

Mark, Jeremy, and Dexter Tiff Roberts. "United States–China Semiconductor Standoff: A Supply Chain under Stress." *Atlantic Council*, 23 Feb. 2023, www.atlanticcouncil.org/in-depth-research-reports/issue-brief/united-states-china-semiconductor-standoff-a-supply-chain-under-stress/.

"NVDA Forecast - Price Target - Prediction for 2025." *TradingView*, 28 Dec. 2024, www.tradingview.com/symbols/NASDAQ-NVDA/forecast/.

Prince, Edemirukewan. "Flag Patterns." *Medium*, Coin Monks, 31 May 2024, medium.com/coinmonks/flag-patterns-9eafb3bdfa54.

"Robotics in Semiconductor Market Size, Share and Growth Report 2032." *Robotics in Semiconductor Market Size, Share and Growth Report 2032*, www.marketresearchfuture.com/reports/robotics-in-semiconductor-market-17764#:~:text=In%202022%2C%20the%20size%20of,(CAGR)%20of%2017.3%25. Accessed 28 Dec. 2024.

Shilov, Anton. "TSMC Posts Q4'23 Earnings: 3nm Revenue Share Jumps to 15%, 5NM Overtakes 7nm for 2023." *AnandTech*, AnandTech, 19 Jan. 2024, www.anandtech.com/show/21239/tsmc-q4-2023-earnings-3nm-revenue-share-jumps-5nm-overtakes-7nm.

Sivasothy, Audrey. "How Is the IOT Impacting the Semiconductor Industry? - Vyrian - Your Primary Source of Semiconductors, Connectors and Leds." *Vyrian*, 17 Aug. 2024, www.vyrian.com/blog/how-is-the-iot-impacting-the-semiconductor-industry/.

Subramanian, Anand. “5 Trending Topics in Semiconductor Technology for the Last Quarter of 2024 and Forward-Looking Insights into 2025.” *TenXer Labs*, 24 Nov. 2024, tenxerlabs.com/resources/blogs/5-trending-topics-in-semiconductor-technology-for-the-last-quarter-of-2024-and-forward-looking-insights-into-2025/.

Swanson, Ana. “U.S. Vies with Allies and Industry to Tighten China Tech Controls.” *The New York Times*, The New York Times, 9 Aug. 2024, www.nytimes.com/2024/08/09/business/economy/china-us-chip-semiconductors.html.

Team Counterpoint. “Global Semiconductor Foundry Market Share: Quarterly.” *Counterpoint*, 26 Nov. 2024, www.counterpointresearch.com/insights/global-semiconductor-foundry-market-share/.

“TSM Stock Price - TSMC Chart.” *TradingView*, 27 Dec. 2024, www.tradingview.com/symbols/NYSE-TSM/.

“TSMC vs. Samsung: The Battle for Semiconductor Dominance.” *ThinkMarkets*, ThinkMarkets, 17 June 2024, www.thinkmarkets.com/en/market-news/tsmc-vs-samsung/.

“U.S. Semiconductors Industry Analysis.” *Simply Wall St*, 28 Dec. 2024, simplywall.st/markets/us/tech/semiconductors.

“Unlocking the Power of Consensus Estimates with Ibes.” *FasterCapital*, fastercapital.com/content/Unlocking-the-Power-of-Consensus-Estimates-with-IBES.html. Accessed 28 Dec. 2024.

Vipond, Tim. “Triangle Patterns - Technical Analysis.” *Corporate Finance Institute*, 15 July 2024, corporatefinanceinstitute.com/resources/career-map/sell-side/capital-markets/triangle-patterns/.

“What Is Rectangle Chart Pattern.” *Investobull*, 30 Apr. 2021, investobull.com/blog/rectangle-chart-pattern/.

Willing, Nicole. “*: Latest Market News: Argus Media.” *Latest Market News*, 30 Sept. 2024, www.argusmedia.com/en/news-and-insights/latest-market-news/2613344-ge-buyers-seek-new-supply-alternatives-as-demand-rises?utm_source=.

Yeh, Esme. “Semiconductor Ecosystem ‘Irreplaceable’: NSTC - Taipei Times.” *TAIPEI TIMES* , 19 July 2024, www.taipeitimes.com/News/taiwan/archives/2024/07/19/2003821024.

Sensors in Stretchable Bioelectronics Based on Nanocomposite By Yujun Sung

Abstract

Nanocomposite-based bioelectronics have developed as an innovative platform in healthcare applications, providing outstanding mechanical compliance, electrical performance, and biocompatibility. This review discuss the integration of 0D, 1D, and 2D nanomaterials into elastomer matrices which allow to overcomes the mechanical mismatch between conventional device materials and soft biological tissues. Optimized percolation networks provide both stretchability of devices while maintaining conductivity under dynamic deformation of organ tissues. Also, improved conductivity and stability of nanocomposite via hybrid crosslinking and surface functionalization of conductive nanofillers is discussed. Next, significant progress in bioelectronic sensors for cardiovascular, brain/peripheral nerve, and muscle through nanocomposite material innovation are examined. Lastly, obstacles that remained in obtaining lasting biostability, scalable fabrication, and multifunctional integration along with clinically viable option to accelerate toward personalized and real-time healthcare monitoring is concluded.

Keyword

Bioelectronics, Nanomaterials, Nanocomposite, Stretchable electronics, Biological tissue interface, Wearable electronics, Implantable devices

1. Introduction

Incorporation of electronic devices into daily life has drastically changed modern life, especially in the field of healthcare.[1,2] Early diagnosis plays a crucial part to improving patient outcomes by providing treatments that may prevent the progression of illness.[3,4] Conventional electronic devices, typically fabricated by bulky and rigid materials such as metals and silicon, experiencing significant obstacles when utilized in healthcare diagnostics. Conventional materials, with elastic moduli exceeding 100 gigapascals (GPa), demonstrate remarkable stiffness compared to biological tissues, which have moduli ranging from approximately 1 kilopascal (kPa) in the brain to 15 kPa in the heart.[5] Their mechanical rigidity represents in sharp contrast to the soft, dynamic properties of biological tissues, resulting in poor adhesion, reduced signal quality, and possible adverse immune responses.[6] The challenges highlight the increasing need for innovative materials and device designs that can seamlessly interact with biological systems, all while ensuring outstanding performance and compatibility with biological tissues. (**Figure 1**)

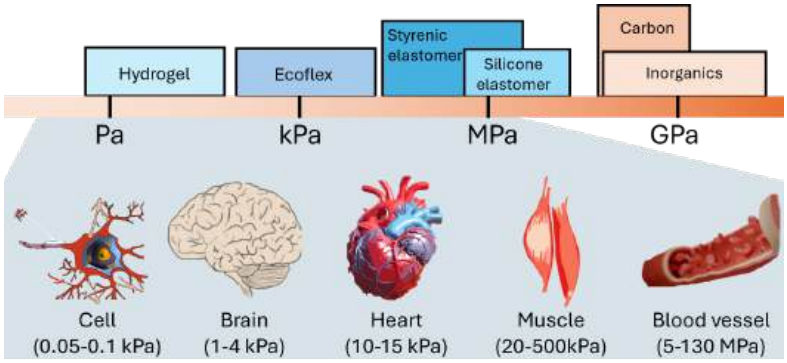


Figure 1. Elastic modulus of electronic materials and biological tissues.

The incorporation of nanoparticles into an elastomer matrix has emerged as an important breakthrough in the development of bio-integrated devices. These novel nanocomposites combine the electrical conductivity of nanoscale fillers with the flexibility and elasticity of elastomers, which makes them distinctly different from previous technologies.[7] Elastomers, composed of intertwined polymer chains, provide the necessary elasticity to match the mechanical properties of human tissues, enabling seamless integration and lasting compatibility.[8] The integration of nanoscale fillers, such as metallic nanowires, nanoparticles, and carbon nanotubes, provides percolation networks that enhance electrical conductivity while preserving mechanical stretchability.

This review provides an investigation of advancements in nanocomposite-based bioelectronics, highlighting their groundbreaking potential in healthcare applications. By reviewing the materials and integration process of zero-dimensional (0D), one-dimensional (1D), and two-dimensional (2D) nanomaterials within elastomer matrices, we discuss the synergistic relationship between mechanical stretchability and electrical performance associated in these mixed structures. This review also studies specific applications in cardiac monitoring, vascular health, neural interfaces, peripheral nerve recovering, and muscle function, emphasizing how nanocomposite overcome persistent obstacles with biointegration, biocompatibility, and device stability. In conclusion, we examine the significant challenges that remain while offering perspectives on future directions in this rapidly evolving field. These insights pave the way for the continued development of next-generation bioelectronic devices for precision medicine and personalized healthcare.

2. Nanocomposite

Nanocomposites are hybrid materials that combine functional nanostructures with elastomeric matrices to achieve exceptional mechanical flexibility and electrical performance. By embedding nanoscale conducting fillers such as nanoparticles, nanowires, or nanosheets into polymers, these composites exhibit properties that surpass those of their individual components. (**Figure 2**) The elastomer matrix provides structural integrity and stretchability, while nanofillers enhance electrical conductivity, optical activity, and thermal stability.[9] Optimizing the type, size, and concentration of fillers, along with tailoring the polymer matrix, enables precise control over

the composite's properties, making nanocomposites essential for applications in wearable sensors,[10] neural interfaces[11], and implantable devices.[12]

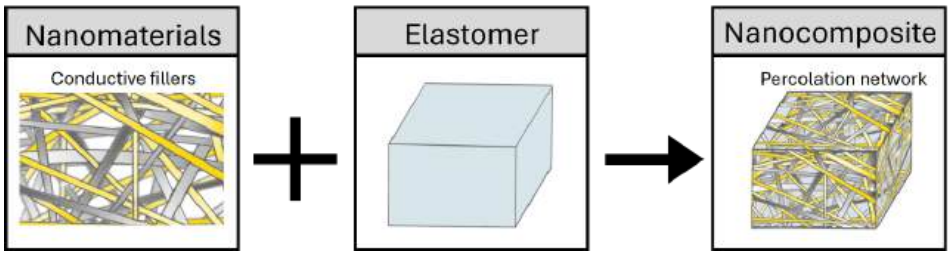


Figure 2. Schematic illustration of nanocomposite fabrication.

Nanocomposites are generally fabricated using either physical blending or in situ polymerization. In physical blending, pre-synthesized nanomaterials is dispersed into elastomer precursors using methods like ultrasonication; in contrast, in situ polymerization synthesizes nanofillers directly within the matrix, which ensure a more uniform distribution and stronger interfacial bonding. Balancing the filler concentration with the matrix crosslinking is quite critical – too much filler can compromise flexibility, and too little might not be sufficient to establish conductive networks. In this chapter, we explore the role of elastomers as matrix materials and the formation of percolation networks, both of which are pivotal for enabling the unique properties of nanocomposites. Overall, these approaches demonstrates significant promise in advancing the field.

2.1 Nanomaterials

Nanomaterials have transformed the field of soft bioelectronics via the fabrication of nanocomposites with exceptional electrical conductivity and softness. These materials, defined by their nanoscale dimensions and unique structural properties, serve as conductive fillers that form percolation networks within elastomer matrices, facilitating charge transport even under mechanical deformation.[7] By integrating nanomaterials into elastomers, researchers have developed stretchable and soft electronic devices that overcome the limitations of conventional, rigid devices.[13] This chapter classifies nanomaterials by dimensions: zero-dimension (0D), one-dimension (1D), and two-dimension (2D). Also, the roles of nanomaterials within nanocomposite along with their dispersion, welding, and stability within the elastomer is discussed. (**Figure 3**)

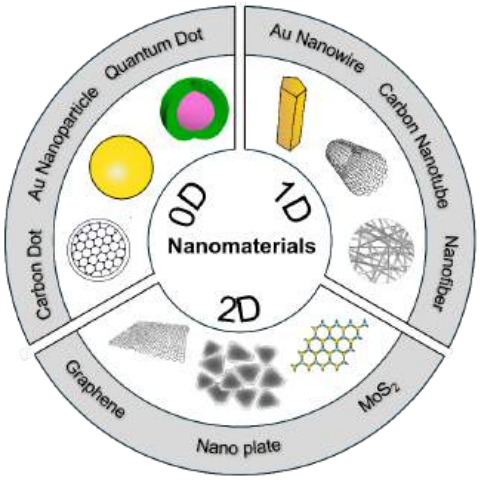


Figure 3. Types of nanomaterials by dimensionality.

2.1.1 Zero-Dimensional (0D) Nanomaterials

Zero-dimensional nanomaterials including nanoparticles, quantum dots (QDs), and carbon dots are distinguished by spatial confinement in all three dimensions. This unique structure endows them with remarkable quantum properties and an exceptionally high surface-area-to-volume ratio, which makes them almost indispensable as conductive fillers in elastomeric matrices for bioelectronic applications.[14] Metallic nanoparticles, such as silver (Ag), gold (Au), and platinum (Pt), are among the most widely studied 0D nanomaterials; they form dense percolation networks that enable efficient charge transport and ensure high conductivity in stretchable composites.

Quantum dots (QDs) are nanoscale semiconductor crystals (1–10 nm) that display tunable optical properties, emitting wavelengths that range from ultraviolet to near-infrared based on their size and composition.[15] Their high quantum yield, photostability, and broad excitation range make them ideal for biosensing and bioelectronics. Surface modifications, such as conjugation with hydrophilic materials or polymers like PEG, enhance water solubility and stability.[16] QDs can function in biosensors either as active probes reacting with analytes to alter fluorescence or as passive labels when conjugated with antibodies. While concerns regarding cadmium toxicity still remains, hybridizing QDs with biocompatible materials helps to improve safety, which in turn enables applications in diagnostics and high-resolution cellular imaging.

Carbon dots (CDs) are 0D nanomaterials with sizes typically below 10 nm, composed of a sp^2/sp^3 carbon framework and abundant surface functional groups such as amine, carboxyl, and hydroxyl.[17] These features provide excellent aqueous solubility, enable functionalization, and make CDs versatile for applications in bioimaging, sensing, and catalysis.[18] CDs present a remarkable properties such as increased photoluminescent quantum yield, controlled excitation and emission, stability toward photobleaching, and biocompatibility. Their distinct property, primarily determined by the level of carbonization, allows significant customization in response to specific application demands. Therefore, this adaptability has established CDs as an innovative material in the field of novel 0D nanomaterials.

2.1.2 One-Dimensional (1D) Nanomaterials

One-dimension nanomaterials, including carbon nanotubes (CNTs), nanowires (NWs), and nanofibers, offer unique advantages for bioelectronics due to their high aspect ratios, which facilitate the formation of efficient percolation networks. These networks require lower filler concentrations compared to 0D materials, preserving the mechanical softness and stretchability of elastomeric composites while maintaining robust electrical conductivity.[19]

Carbon nanotubes are among the most extensively studied 1D nanomaterials, known for their exceptional electrical conductivity, mechanical strength, and thermal stability.[20] Single-walled carbon nanotubes (SWCNTs) exhibit either metallic or semiconducting properties depending on their chirality, while multi-walled carbon nanotubes (MWCNTs) provide high conductivity.[21] However, their agglomeration tendencies necessitate dispersion techniques such as chemical functionalization and surfactant-assisted processing, which enhance their uniform distribution and interfacial bonding within elastomer matrices.[22]

Metallic nanowires, such as silver (Ag), gold (Au), and copper (Cu) nanowires, are another class of 1D nanomaterials with widespread applications in soft electronics. The elongated geometry of nanowires facilitates the creation of conductive networks with low contact resistance, even under mechanical strain. Optimizing the length and diameter of nanowires has been shown to improve their packing density and network connectivity, enhancing the conductivity and mechanical robustness of the resulting composites.[23] For example, Ag nanowire-based composites are widely used in applications such as flexible displays, wearable sensors, and stretchable electrodes for medical devices.

2.1.3 Two-Dimensional (2D) Nanomaterials

Two-dimension nanomaterials, such as graphene and transition metal dichalcogenides (TMDs), are distinguished by their planar structures, which impart exceptional electrical, mechanical, and optical properties. These features make them highly attractive for the development of soft bioelectronics, particularly in applications demanding high conductivity and mechanical robustness.

Graphene, a monolayer of sp^2-hybridized carbon atoms, is renowned for its remarkable electrical conductivity, mechanical strength, and optical transparency. Its high Young's modulus (~1 TPa) and flexibility make it an ideal candidate for stretchable and transparent bioelectronic devices. Reduced graphene oxide (rGO), a chemically modified form of graphene, has gained popularity for its improved dispersibility in polymer matrices, enabling the formation of homogeneous composites with enhanced electrical and mechanical performance.

Transition metal dichalcogenides, such as molybdenum disulfide (MoS_2), represent another class of 2D nanomaterials with significant potential for bioelectronics.[24] MoS_2, with its direct bandgap and high quantum efficiency, is particularly suited for optoelectronic applications, including photodetectors and energy storage systems.[25] Combining TMDs with graphene has led to the development of heterostructures with synergistic properties, enabling advanced functionalities in soft electronic devices.

Ultrathin monocrystalline nanomembranes, including silicon (Si) and gallium nitride (GaN), have also emerged as promising candidates for high-performance bioelectronics. Fabricated using advanced top-down approaches, these nanomembranes offer superior flexibility and electronic performance, making them integral to devices requiring high-resolution sensing and signal processing.[26] The versatility of 2D nanomaterials continues to drive innovations in bioelectronics, enabling applications that combine exceptional conductivity, flexibility, and multifunctionality.

2.2 Elastomers as Matrix Materials

Elastomers form the foundational matrix for nanocomposites, offering the softness, stretchability, and mechanical resilience required for biointegrated applications. These elastic polymers consist of long polymer chains connected by crosslinks, which allow them to deform under stress and return to their original shape once the stress is removed. Elastomers are particularly suited for interfacing with soft biological tissues, such as skin and the heart, which undergo dynamic strains. The mechanical properties of elastomers, including elasticity, toughness, and Young's modulus, can be finely tuned by altering the crosslinking type and density.

Crosslinking in elastomers can be classified as physical, chemical, or a hybrid of the two. Physical crosslinking involves reversible interactions, such as hydrogen bonds, ionic interactions, or π–π stacking, which allow elastomers to be reprocessed and adapted to environmental changes. For instance, styrenic block copolymers like poly(styrene-butadiene-styrene) (SBS) exhibit dual-phase structures, where soft and hard domains coexist to provide both elasticity and strength.[27] Chemical crosslinking, by contrast, relies on covalent bonds that confer superior stability and durability. A representative example is poly(dimethylsiloxane) (PDMS), a chemically crosslinked elastomer widely used in biomedical applications for its biocompatibility and robustness. Hybrid crosslinking systems combine the strengths of both approaches, achieving reversible adaptability from physical interactions and permanent mechanical stability from chemical bonds. These hybrids are particularly beneficial in dynamic environments, such as wearable and implantable bioelectronics, where tunable mechanical and functional properties are crucial.

While elastomers dominate the landscape of nanocomposite matrices, hydrogels represent an alternative for specific applications requiring soft, tissue-mimicking properties. Hydrogels, being hydrophilic networks capable of absorbing large amounts of water, are often employed in drug delivery systems, skin-mounted sensors, and neural interfaces. However, their lower mechanical durability compared to elastomers makes them more suitable for niche applications rather than as a primary matrix material.

2.3 Optimization of Percolation Network

The percolation network, formed by interconnected nanofillers within a polymer matrix, underpins the electrical performance of conductive nanocomposites. High-aspect-ratio fillers

such as carbon nanotubes (CNTs) and silver nanowires (AgNWs) efficiently form conductive pathways at lower concentrations compared to spherical fillers. For example, the elongated geometry of AgNWs facilitates robust connectivity even under deformation. Surface functionalization techniques, such as polymer coatings or ligand exchange, enhance dispersion and compatibility, ensuring reliable conductivity.[28]

Maintaining network integrity under mechanical strain is a critical challenge. Hybrid fillers, combining materials like CNTs and AgNWs, create additional pathways to sustain conductivity, while techniques like thermal and optical welding strengthen filler junctions. These strategies ensure the durability and functionality of nanocomposites, making them ideal for dynamic bioelectronic applications.[29]

Conductive networks rely on the homogeneous dispersion and interconnection of fillers, achieved through surface functionalization and welding techniques. Functionalization of nanomaterials with hydrophobic ligands, such as hexylamine or polyethylene glycol (PEG), reduce aggregation and contact resistance, thus improving their compatibility with elastomer matrices. Welding methods, such as plasmonic welding and Joule heating, strengthen junctions between conductive fillers, enhancing electron transport and mechanical durability.[30,31] Hybrid filler systems, combining materials like silver nanowires (AgNWs) with carbon nanotubes (CNTs), provide additional conductive pathways, ensuring both high conductivity and mechanical robustness under repeated strain.

3. Sensors based on Nanocomposite

Nanocomposite-based sensors combine the stretchability of elastomeric matrices with the exceptional electrical properties of conducting nanomaterials, enabling seamless integration with dynamic biological tissues. These sensors address challenges associated with traditional rigid devices, such as mechanical mismatch and poor biocompatibility, making them ideal for advanced diagnostic and therapeutic applications. This chapter explores the latest innovations in nanocomposite-based sensors, focusing on their applications in the heart, blood vessels, brain, peripheral nerves, and muscles. (**Figure 4**)

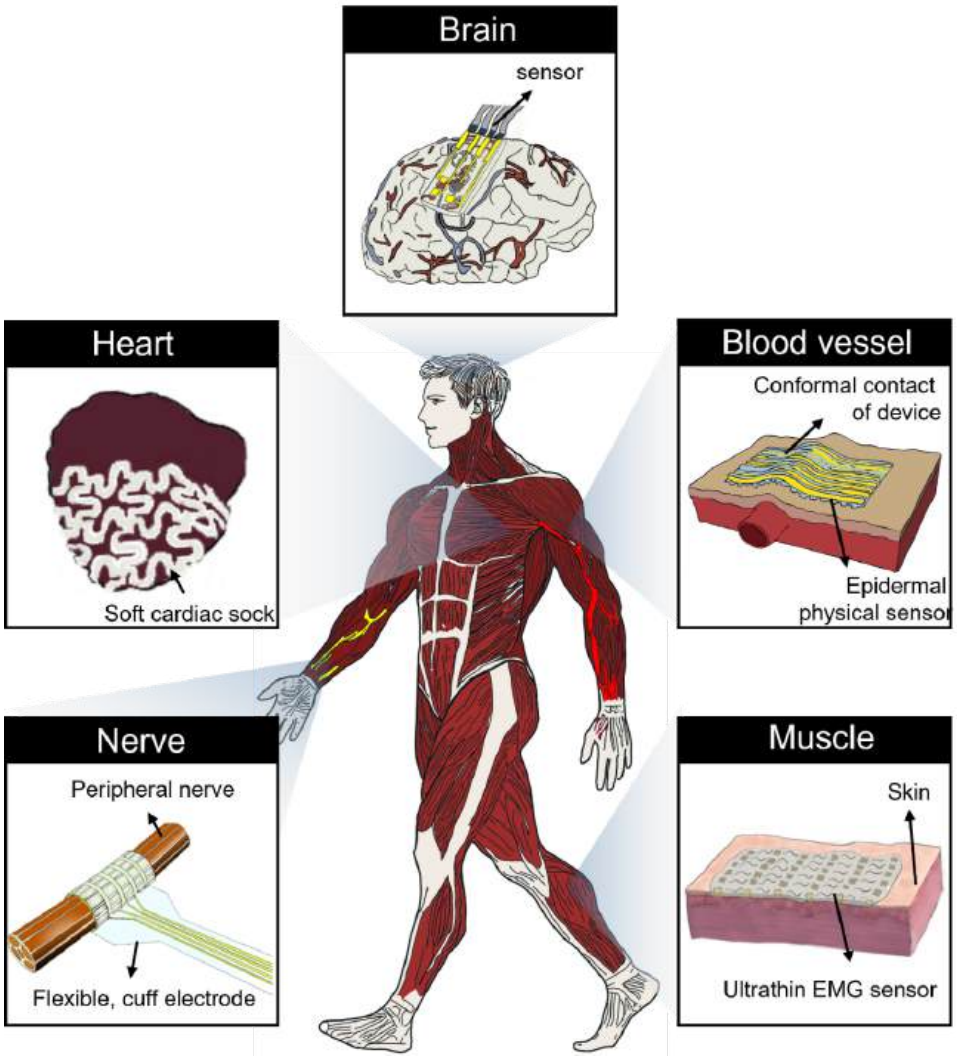


Figure 4. Application of nanocomposite-based sensors as wearable and implantable bioelectronics.

3.1. Heart

Nanocomposite-based sensors have enabled transformative progress in cardiac monitoring and therapy by addressing the limitations of traditional rigid electrodes. The unique combination of stretchability, mechanical compliance, and electrical conductivity in nanocomposites allows epicardial devices to conform seamlessly to the heart's dynamic and curved surfaces. This level of biointegration is critical for capturing high-resolution electrophysiological (EP) signals and delivering effective therapeutic interventions, paving the way for improved cardiac diagnostics and treatment.

The rhythmic contraction and relaxation of the myocardium, governed by EP signals, are essential for maintaining efficient blood circulation. Conventional ECG electrodes, composed of rigid metals and gel, often suffer from performance degradation due to gel drying or motion artifacts. To overcome these issues, a conductive dry adhesive (CDA) sensor was developed using a carbon-based PDMS nanocomposite reinforced with one-dimensional carbon nanotubes

(CNTs) and two-dimensional graphene flakes. This structure formed a conductive percolation network within the elastomer matrix, with additional enhancements provided by graphite and graphene nanopowders. Mushroom-shaped micropillar arrays inspired by gecko adhesion enabled robust skin attachment under dynamic conditions. The CDA sensor demonstrated superior functionality, capturing ECG signals comparable to commercial electrodes while maintaining stable performance during underwater operations and vigorous motion, where conventional electrodes failed.

Park et al. advanced epicardial devices by developing a nanocomposite mesh electrode using ligand-exchanged silver nanowires (Ag NWs) embedded in a styrene-butadiene-styrene (SBS) matrix.[12] This elastic mesh structure conformed seamlessly to the heart's surface, ensuring mechanical compliance while delivering large-area electrical stimulation to support cardiac contractility. The electrode exhibited high conductivity and stretchability, enabling precise ECG recording and effective therapeutic intervention. Building on this innovation, Choi et al. introduced a nanocomposite combining silver-gold core–sheath nanowires (Ag–Au NWs) with SBS rubber, significantly improving the device's biocompatibility and durability by protecting the silver core from oxidation.[32] The phase-separated microstructure of this composite, achieved through room-temperature drying, enhanced its stretchability and conductivity, enabling stable ECG recordings with a high signal-to-noise ratio and reliable mechanical support for dynamic cardiac movements.

These advancements in nanocomposite-based sensors have redefined the landscape of cardiac monitoring and therapy, providing highly conformal, durable, and biocompatible solutions for capturing precise EP signals and delivering targeted electrical stimulation. By bridging the gap between mechanical compliance and electrical functionality, these innovations offer a transformative approach to improving cardiac healthcare.

3.2 Blood vessel

Biophysical signals derived from vascular dynamics, such as pulse waveforms, pulse velocity, and blood pressure (BP), are critical indicators for diagnosing cardiovascular conditions like arrhythmias and hypertension. While cuff-based sphygmomanometers are widely used in clinical practice, they are unsuitable for continuous monitoring, limiting their ability to capture real-time, dynamic changes. To overcome these limitations, soft bioelectronics have emerged as a transformative technology, offering skin-interfaced systems equipped with resistive, capacitive, piezoelectric, and optical sensors. These advanced sensing modalities provide high sensitivity, rapid response times, and minimal hysteresis, enabling precise and continuous monitoring of vascular signals under subtle strains induced by blood flow.

Piezoresistive sensors have become a cornerstone of wearable vascular monitoring due to their ease of fabrication and high sensitivity. Du et al. fabricated a piezoresistive sensor by growing graphene on a non-woven fabric substrate.[33] This simple yet effective method produced a wearable, cloth-compatible sensor with excellent sensitivity of approximately 0.057 kPa^{-1} and a gauge factor (GF) of −7.1 under 1% strain, maintaining functionality over 10,000 cycles. The

graphene-based structure ensured robust conductivity while allowing for mechanical flexibility.

Ai et al. developed a piezoresistive sensor using reduced graphene oxide (rGO) embedded between PDMS films.[34] The fabrication process involved layer-by-layer assembly to ensure uniform rGO distribution, optimizing its piezoresistive properties. This device, when tested on a healthy subject, successfully captured detailed pulse waveforms, including the diastolic tail, which traditional tonometry methods failed to detect. With a sensitivity of 50.9 kPa^{-1}, a limit of detection (LOD) of 3 Pa, and minimal power consumption (~1 μW), the sensor demonstrated exceptional potential for continuous BP monitoring. Its 50 ms response time and high durability made it suitable for long-term wearable applications.

Wu et al. utilized a laser scribing technique to fabricate a graphene-based piezoresistive sensor for non-invasive BP monitoring.[35] This method involved precision patterning of graphene on a flexible substrate, enabling an average sensitivity of 12.3 kPa^{-1}. The sensor's design allowed BP measurements at multiple body locations, providing versatility for wearable healthcare systems. Luo et al. further advanced this field by integrating graphene nanowalls, PDMS, and zinc oxide (ZnO) into a microfabricated piezocapacitive sensor.[36] The device demonstrated a sensitivity of 22.3 kPa^{-1}, a response time of 25 ms, and a sensing range of 22 kPa, highlighting its suitability for wearable pulse monitoring applications.

Soft bioelectronics have revolutionized vascular monitoring by addressing the limitations of traditional BP measurement techniques, enabling continuous and precise detection of biophysical signals. Graphene-based technologies, including piezoresistive and piezocapacitive sensors, have demonstrated significant advancements in sensitivity, durability, and adaptability. Innovations such as rGO-PDMS sensors, with high sensitivity and rapid response times, and laser-scribed graphene designs for versatile body placement, highlight their potential for long-term wearable healthcare applications. These advancements position soft bioelectronics as critical tools in the development of non-invasive, real-time monitoring systems for cardiovascular health.

3.3 Brain

Nanocomposite-based sensors have redefined neuroengineering by addressing the challenges posed by the brain's delicate structure and sensitivity to inflammation. The development of bioelectronics for the brain requires materials that are mechanically soft, chemically biocompatible, and capable of forming stable interfaces with neural tissue. Unlike conventional materials, nanocomposites provide an unparalleled combination of flexibility, conductivity, and biocompatibility, enabling minimally invasive, long-term monitoring and modulation of brain activity. These advances hold the potential to revolutionize applications ranging from neural diagnostics to brain-computer interfaces.

The monitoring of brain activity using electroencephalography (EEG) can be classified into two primary methods: invasive intracranial EEG (iEEG) and noninvasive surface EEG (sEEG). Devices for both methods demand mechanical softness and non-toxicity, particularly for iEEG applications. Tybrandt et al. addressed these requirements by creating flexible neural grid

electrodes using PDMS and Au–TiO_2 NWs.[37] The electrodes, encapsulated within a PDMS layer, demonstrated an initial conductivity of approximately 16,000 S cm^{-1}. Although the resistance increased under strain, the device successfully recorded somatosensory evoked potentials (SSEPs) from the somatosensory cortex of rats. The electrodes spontaneously conformed to the brain's pial surface after implantation, maintaining low resistance and high signal quality for over three months.

In another example, canal-type ear electrodes (CEE) composed of a multiwalled carbon nanotube (MWCNT)/PDMS nanocomposite were developed for noninvasive EEG monitoring.[38] Resembling the soft cover of a canal-style earphone, the device exhibited a low Young's modulus (~1 MPa), enabling painless insertion into the ear. The MWCNT/PDMS CEE demonstrated excellent performance in capturing alpha rhythm waveforms, comparable to traditional on-scalp electrodes, while offering superior comfort and stability. The low impedance of the device (~1 MΩ) contributed to its ability to record various EEG signals, including steady-state visually evoked potentials and auditory evoked potentials, without causing irritation or toxicity, as confirmed by an alive/dead assay.

Nanocomposites also extend beyond electrical sensing, enabling advanced functionalities for neurological applications. Conductive nanoparticles, such as gold, silver, and platinum, reduce impedance and enhance charge injection, critical for deep-brain signal recording. For example, 3D-printed silver nanoparticles coated with gold have been used to fabricate high-density multichannel electrodes for precise neural recordings. Platinum nanoclusters further improve charge transfer and enable high-resolution recordings on microelectrodes. Organic nanomaterials, such as graphene, have been integrated into flexible field-effect transistor (FET)-based probes, leveraging their high electrical mobility and biocompatibility for neural stimulation and recording. Graphene electrodes, free from MRI distortion, offer significant advantages for multimodal imaging and recording applications.

Beyond electrical applications, nanocomposite bioelectronics enable optical and chemical sensing functionalities. Gold-decorated TiO_2 nanowire arrays have been used in retinal implants to restore vision in blind mice by mimicking damaged photoreceptors. Stretchable neurotransmitter sensors, incorporating catalytic metal oxide nanoparticles and graphene nanofiber networks, enable real-time neurochemical monitoring. These innovations underscore the transformative potential of nanocomposite-based bioelectronics in neuroengineering, addressing the complex requirements of brain applications and paving the way for breakthroughs in neural therapeutics and brain-computer interfaces.

3.4 Peripheral nerve

Peripheral nerve interfaces are essential for restoring sensory and motor functions, including pain relief, walking rehabilitation, and prosthetic control. Peripheral nerves, composed of motor and sensory axons bundled with Schwann cells, blood vessels, and connective tissue, exhibit high flexibility (Young's modulus ≈ 0.5-10 MPa) and constant movement.[39] Designing electrodes for these interfaces requires softness, durability, and stable electrical performance to

minimize nerve atrophy and ensure long-term functionality. Stiffness mismatches between electrodes and nerves can lead to glial encapsulation, nerve degeneration, and reduced stimulation efficiency. Traditional electrodes, made from rigid materials like platinum (≈200 GPa) and silicone rubber (≈2 MPa), lack the flexibility and biocompatibility required for long-term implantation.[6] Nanocomposites incorporating materials like gold-coated silver nanowires, platinum black, PEDOT:PSS-coated silver nanoparticles, and carbon nanotubes (CNTs) address these limitations, combining high conductivity, charge transfer capacity, and mechanical compliance.

Heo et al. demonstrated a flexible electrode fabricated by inkjet-printing silver nanoparticles (AgNPs) onto an electrospun polyimide (PI) nanofiber membrane (~136 μm thick). [40] Heat treatment at 140°C fused the AgNPs onto the nanofibers, followed by electrochemical polymerization of PEDOT:PSS, significantly reducing impedance to ~123 Ohms at 1 kHz, 7.5 times lower than traditional cuff electrodes. This electrode maintained stable nerve recordings for up to 12 weeks in rat sciatic nerves, showcasing its potential for long-term use. Similarly, Seo et al. developed a self-healing neuroprosthetic device using gold nanoshell-coated silver flakes (AuNS-AgF) embedded in a self-healing polymer (SHP) matrix.[41] The SHP leveraged dynamically recross-linking hydrogen bonds to repair material degradation caused by repetitive nerve deformations. Implanted on rat sciatic nerves, the device enabled stable electrical modulation, eliciting muscle contractions and accurately tracking joint movements during treadmill walking.

Advancements in peripheral nerve interfaces leverage nanocomposites and innovative designs to address the unique demands of nerve flexibility, biocompatibility, and durability. From electrodes with low impedance and high mechanical compliance to self-healing and bioresorbable systems, these technologies offer tailored solutions for nerve repair and regeneration. Future research should focus on refining these approaches for clinical translation, enabling broader accessibility and improved outcomes in peripheral nerve applications.

3.5 Muscle

Electromyography (EMG) sensors are indispensable tools for assessing muscle function, with applications spanning brain-machine interfaces and rehabilitation technologies. Despite their utility, traditional EMG systems often face challenges such as high impedance and low signal-to-noise ratios (SNR), primarily due to inadequate skin-electrode contact. To overcome these limitations, recent innovations have leveraged flexible nanocomposites and functional elastomers, enabling conformal skin integration, gel-free operation, and high-fidelity signal acquisition. These advancements mark a significant step toward improving the usability and reliability of EMG systems in wearable healthcare devices.

Graphene elastomeric electrodes (GETs) have emerged as a groundbreaking solution for EMG sensing.[42] Synthesized via chemical vapor deposition (CVD) and patterned with a mechanical cutter plotter, GETs demonstrated exceptional flexibility and skin conformability when applied to the forearm. The dry, gel-free GETs achieved signal quality comparable to

traditional Ag/AgCl gel electrodes, making them a promising non-invasive alternative for wearable applications. Similarly, silver nanowire (AgNW)/PDMS composites offer a compelling combination of electrical conductivity (~5000 S cm^{-1}) and mechanical compliance.[43] Integrated onto the wrist using a velcro strap, these electrodes recorded EMG signals with an SNR of 24.7 dB, closely matching the 27.3 dB of conventional electrodes, all while eliminating the need for conductive gels.

In another innovative approach, Matsuhisa et al. developed a wearable electrode array using Ag-flake-based elastic conductor ink printed onto stretchable fabric substrates. [44] This fabrication method, relying on surfactant-induced phase separation, produced nanocomposites with superior conductivity and stretchability. The electrode array conformally adhered to the skin, accurately recording EMG signals. Noise reduction was achieved with a low-pass filter, while an organic amplifier enhanced signal strength, enabling precise discrimination of hand motions. This system demonstrated the potential of nanocomposites for advanced muscle function analysis in healthcare applications.

The integration of advanced nanocomposites such as GETs, AgNW/PDMS composites, and Ag-flake elastic conductor inks has redefined wearable EMG technology. By addressing key limitations of traditional systems, these innovations provide high SNR, enhanced skin conformity, and gel-free operation, paving the way for user-friendly, high-performance systems tailored to healthcare and rehabilitation. These breakthroughs not only enhance diagnostic and therapeutic capabilities but also signify a major leap toward the widespread adoption of wearable EMG systems in clinical and everyday settings.

4. Conclusion

Nanocomposites have redefined the landscape of soft bioelectronics, offering unparalleled opportunities to bridge the gap between rigid electronic systems and the dynamic, soft nature of biological tissues. By integrating nanoscale conductive fillers into elastomeric matrices, these materials achieve a unique balance of stretchability, biocompatibility, and electrical performance, enabling transformative applications in diagnostics, monitoring, and therapy.

The development of advanced nanomaterials, including 0D nanoparticles and quantum dots, 1D nanowires and nanotubes, and 2D materials like graphene, has driven significant progress in the field. These materials have been harnessed to create high-performance bioelectronic devices for cardiac health, vascular monitoring, neural interfaces, peripheral nerve repair, and muscle function assessment. The ability to form percolation networks within soft matrices has enabled these systems to maintain conductivity under mechanical deformation, while innovations in hybrid crosslinking and surface functionalization have enhanced their mechanical durability and long-term stability.

Despite these advancements, challenges remain. Ensuring scalability, minimizing cytotoxicity, and achieving stable biointerfaces for long-term implantation are critical areas for future research. Additionally, the integration of multifunctional capabilities, such as self-healing,

biodegradability, and wireless operation, will be essential to fully realize the potential of these materials in clinical and wearable applications.

Looking forward, the synergy between materials science, bioengineering, and healthcare innovation will continue to drive breakthroughs in nanocomposite-based bioelectronics. By addressing existing limitations and expanding the range of applications, these technologies hold the promise to revolutionize healthcare, enabling real-time, non-invasive, and personalized solutions for a wide array of medical challenges.

Acknowledgement

I would like to thank Calvin Cho for his guidance and encouragement during the process of this review.

Works Cited

1.Sunwoo, S. H., Ha, K. H., Lee, S., Lu, N. & Kim, D. H. Wearable and Implantable Soft Bioelectronics: Device Designs and Material Strategies. *Annu Rev Chem Biomol Eng* **12**, 359–391 (2021).

2.Chitrakar, C., Hedrick, E., Adegoke, L. & Ecker, M. Flexible and Stretchable Bioelectronics. *Materials* **15**, (2022).

3.Ahmed, F. Z. *et al.* Early diagnosis of cardiac implantable electronic device generator pocket infection using 18F-FDG-PET/CT. *Eur Heart J Cardiovasc Imaging* **16**, 521–530 (2015).

4.Keum, D. H. *et al.* Wireless smart contact lens for diabetic diagnosis and therapy. *Sci Adv* **6**, eaba3252 (2020).

5.Shim, H. J., Sunwoo, S., Kim, Y., Koo, J. H. & Kim, D. Functionalized Elastomers for Intrinsically Soft and Biointegrated Electronics. *Adv Healthc Mater* **10**, 2002105 (2021).

6.Park, J., Lee, Y., Kim, T. Y., Hwang, S. & Seo, J. Functional Bioelectronic Materials for Long-Term Biocompatibility and Functionality. *ACS Appl Electron Mater* **4**, 1449–1468 (2022).

7.Choi, S., Han, S. I., Kim, D.-H. H. D. D. H., Hyeon, T. & Kim, D.-H. H. D. D. H. High-performance stretchable conductive nanocomposites: Materials, processes, and device applications. *Chem Soc Rev* **48**, 1566–1595 (2019).

8.Cho, K. W. *et al.* Soft Bioelectronics Based on Nanomaterials. *Chem Rev* **122**, (2022).

9.Shim, H. J., Sunwoo, S. H., Kim, Y., Koo, J. H. & Kim, D. H. Functionalized Elastomers for Intrinsically Soft and Biointegrated Electronics. *Adv Healthc Mater* **10**, 1–33 (2021).

10.Fondjo, F., Lee, D. S., Howe, C., Yeo, W.-H. & Kim, J.-H. Synthesis of a Soft Nanocomposite for Flexible, Wearable Bioelectronics. in *2017 IEEE 67th Electronic Components and Technology Conference (ECTC)* 780–785 (IEEE, 2017). doi:10.1109/ECTC.2017.195.

11.Sunwoo, S. H. *et al.* Advances in Soft Bioelectronics for Brain Research and Clinical Neuroengineering. *Matter* **3**, 1923–1947 (2020).

12.Park, J. *et al.* Electromechanical cardioplasty using a wrapped elasto-conductive epicardial mesh. *Sci Transl Med* **8**, 344ra86 (2016).

13.Joo, H., Jung, D., Sunwoo, S. H., Koo, J. H. & Kim, D. H. Material Design and Fabrication Strategies for Stretchable Metallic Nanocomposites. *Small* **16**, 1–19 (2020).

14.Wang, Z., Hu, T., Liang, R. & Wei, M. Application of Zero-Dimensional Nanomaterials in Biosensing. *Front Chem* **8**, 320 (2020).

15.Baskoutas, S. & Terzis, A. F. Size-dependent band gap of colloidal quantum dots. *J Appl Phys* **99**, 013708 (2006).

16.Kim, T.-H. *et al.* Fully Stretchable Optoelectronic Sensors Based on Colloidal Quantum Dots for Sensing Photoplethysmographic Signals. *ACS Nano* **11**, 5992–6003 (2017).

17.Lim, S. Y., Shen, W. & Gao, Z. Carbon quantum dots and their applications. *Chem Soc Rev* **44**, 362–381 (2015).

18.Won, H. J. *et al.* Diselenide-Bridged Carbon-Dot-Mediated Self-Healing, Conductive, and Adhesive Wireless Hydrogel Sensors for Label-Free Breast Cancer Detection. *ACS Nano*

14, 8409–8420 (2020).
19. Garnett, E., Mai, L. & Yang, P. Introduction: 1D Nanomaterials/Nanowires. *Chem Rev* **119**, 8955–8957 (2019).
20. Baddour, C. E. & Briens, C. Carbon Nanotube Synthesis: A Review. *International Journal of Chemical Reactor Engineering* **3**, R3 (2005).
21. Yellampalli, S. *Carbon Nanotubes - Synthesis, Characterization, Applications*. (InTech, 2011). doi:10.5772/978.
22. Hu, N. *et al.* The electrical properties of polymer nanocomposites with carbon nanotube fillers. *Nanotechnology* **19**, 215701 (2008).
23. Liu, B., Luo, Z., Zhang, W., Tu, Q. & Jin, X. Silver nanowire-composite electrodes for long-term electrocardiogram measurements. *Sens Actuators A Phys* **247**, 459–464 (2016).
24. Choi, C. *et al.* Curved neuromorphic image sensor array using a MoS2-organic heterostructure inspired by the human visual recognition system. *Nat Commun* **11**, 5934 (2020).
25. Chen, X. *et al.* CVD-grown monolayer MoS2 in bioabsorbable electronics and biosensors. *Nat Commun* **9**, 1690 (2018).
26. Rogers, J. A., Lagally, M. G. & Nuzzo, R. G. Synthesis, assembly and applications of semiconductor nanomembranes. *Nature* **477**, 45–53 (2011).
27. Liu, Y. *et al.* Effect of physical and chemical crosslinking structure on fatigue behavior of styrene butadiene elastomer. *J Appl Polym Sci* **131**, 40917 (2014).
28. Liang, J. *et al.* Silver Nanowire Percolation Network Soldered with Graphene Oxide at Room Temperature and Its Application for Fully Stretchable Polymer Light-Emitting Diodes. *ACS Nano* **8**, 1590–1600 (2014).
29. Li, J. *et al.* Correlations between Percolation Threshold, Dispersion State, and Aspect Ratio of Carbon Nanotubes. *Adv Funct Mater* **17**, 3207–3215 (2007).
30. Oh, J. S., Oh, J. S., Kim, T. H. & Yeom, G. Y. Efficient metallic nanowire welding using the Eddy current method. *Nanotechnology* **30**, 065708 (2019).
31. Garnett, E. C. *et al.* Self-limited plasmonic welding of silver nanowire junctions. *Nat Mater* **11**, 241–249 (2012).
32. Choi, S. *et al.* Highly conductive, stretchable and biocompatible Ag–Au core–sheath nanowire composite for wearable and implantable bioelectronics. *Nat Nanotechnol* **13**, 1048–1056 (2018).
33. Du, D., Li, P. & Ouyang, J. Graphene coated nonwoven fabrics as wearable sensors. *J Mater Chem C Mater* **4**, (2016).
34. Ai, Y. *et al.* An ultrasensitive flexible pressure sensor for multimodal wearable electronic skins based on large-scale polystyrene ball@reduced graphene-oxide core-shell nanoparticles. *J Mater Chem C Mater* **6**, (2018).
35. Wu, Q. *et al.* Triode-Mimicking Graphene Pressure Sensor with Positive Resistance Variation for Physiology and Motion Monitoring. *ACS Nano* **14**, (2020).
36. Luo, S. *et al.* Microconformal electrode-dielectric integration for flexible ultrasensitive

robotic tactile sensing. *Nano Energy* **80**, (2021).
37. Tybrandt, K. *et al.* High-Density Stretchable Electrode Grids for Chronic Neural Recording. *Advanced Materials* **30**, (2018).
38. Hoon Lee, J. *et al.* CNT/PDMS-based canal-typed ear electrodes for inconspicuous EEG recording. *J Neural Eng* **11**, 046014 (2014).
39. Rosso, G. & Guck, J. Mechanical changes of peripheral nerve tissue microenvironment and their structural basis during development. *APL Bioeng* **3**, (2019).
40. Heo, D. N. *et al.* Flexible and Highly Biocompatible Nanofiber-Based Electrodes for Neural Surface Interfacing. *ACS Nano* **11**, (2017).
41. Seo, H. *et al.* Durable and Fatigue‐Resistant Soft Peripheral Neuroprosthetics for In Vivo Bidirectional Signaling. *Advanced Materials* **33**, 2007346 (2021).
42. Cheng, L., Li, J., Guo, A. & Zhang, J. Recent advances in flexible noninvasive electrodes for surface electromyography acquisition. *npj Flexible Electronics* vol. 7 Preprint at https://doi.org/10.1038/s41528-023-00273-0 (2023).
43. Zhou, Z., Wang, H., Zhu, Z., Yang, H. & Zhang, Q. Enhanced dielectric, electromechanical and hydrophobic behaviors of core-shell AgNWs@SiO2/PDMS composites. *Colloids Surf A Physicochem Eng Asp* **563**, 59–67 (2019).
44. Matsuhisa, N. *et al.* Printable elastic conductors with a high conductivity for electronic textile applications. *Nat Commun* **6**, (2015).

Integration of AI Into Our Society: Opportunities and Challenges By Aarav Gupta

Keywords

Rational architecture, degrees of personhood, AI and society, human vs. artificial intelligence, fairness in AI decision-making, practical and theoretical reasoning in AI

Abstract

From the 1940s to the 2010s, Artificial Intelligence (AI) has undoubtedly played a substantial role in shaping the world, yet bias and unfairness in AI systems remain both longstanding and seemingly intractable challenges. Static measures of fairness employed during training do not adequately resolve biases that spontaneously present themselves in the dynamic, real-world scenarios where social dynamics and demographic variables change over time. Such limitations can thus result in biased or advantages/disadvantages for specific sections of people making all the conclusions and decisions using AI systems that are always untrustworthy or biased.

Section 1. Introduction

- **History of AI and how do human and artificial intelligence compare**
- **Integration of AI into our society, challenges, opportunities/solutions**
- **Regulations**

Knowing the differences between human and machine intelligence tells a little about what holds Artificial Intelligence (AI) back. Human intelligence is flexible and tailored by experience, thus people adapt and learn across the scope of lots of different environments. In contrast, AI models are typically narrow in their scope of application and often require really big datasets to actually function. This is more evident in high-stakes areas such as using AI in order to decide on creditworthiness, hiring decisions, or predicting recidivism. In these domains, the shortcomings of the AI system concerning emotional intelligence, creativity, and subtle reasoning also risk perpetuating existing inequalities and sustaining discriminatory practices.

The interaction of AI and human decisions reveals a host of complicated ethical problems, especially with regard to bias and equity. While human reasoning in many instances embeds emotional and moral insight into its working process, AI operates solely based on data that is fed to it, making it very susceptible to the existing biases within the data. Some promising solutions are real-time bias auditing and fairness constraints during training, but most of them come with a catch, such as reduced accuracy, and that is quite telling of the challenges in balancing fairness with performance.

This paper looks at different concepts discussed by philosopher John Pollock regarding the use of personhood and rationality within AI systems. It is a question of whether AI might someday be seen to approach a form of personhood that raises questions about what it means to

be "intelligent" or "rational" and whether AI systems, in their ever-growing complexity, could ever be more than simple tools. It is these philosophical ideas that really engage a deeper conversation about AI ethics and perhaps suggest that intelligence might be defined by rational behavior rather than qualities that are simply similar to those of humans.

If this is not enough, the regulatory landscape introduces yet another layer of complexity. From European data protection laws to specific U.S. regulations concerning AI in hiring, efforts are being made around the world to handle issues related to transparency, accountability, and ethical deployment of AI. The catch, however, is that for the most part, existing regulations grossly fall short of covering the full spectrum of issues that have cropped up with AI's pervasive role in decision-making. The intent of this paper is to point out the gaps in existing frameworks and also presents a notion of other policies that might give a better guard to ethical standards and elicit more care for individual rights.

In the Cognitive Computation Article, Abbass (2019) outlines the journey of AI from simple cognitive augmentation tools to autonomous agents capable of enhanced decision-making. He states that the trust of human-AI relationships is somewhat indispensable, and that both static and adaptive function allocation methodologies will play important roles in defining these interactions. While the static approaches, including economic or comparison-based allocations, often fail to consider dynamic environments, adaptive methodologies such as CoCyS introduce autonomous relationship managers-e-cookies-to increase the level of trust and cooperation between humans and AI systems. These developments reflect the growing ability of AI to cope with complex environments and pose new challenges in the domain of risk management and governance [27].

In contrast, Durt (2022) and Bellaiche et al. (2023) discuss the integration of AI into society and culture, providing insight into its impact beyond functional performance. Durt examines the traditional object-subject dichotomy in AI, suggesting that AI integrates into the human lifeworld in a unique way, by way of navigating and transforming meaning within social and cultural structures. This places AI in an active instead of passive role in shaping how humans experience the world, and for which more nuanced governance frameworks are needed [28].

Bellaiche et al. extend this perspective to the study of public biases against AI-generated artifacts. Their findings show that AI is often perceived as lacking human effort and authenticity in its creativity, even when outputs are of comparable quality. These biases underlie some deep cultural values about human agency and acceptance of AI technologies in creative and social domains [29].

Section 2: History of AI

The aim of this section is to cover the main points of the History of AI, and it will start from its theoretical underpinnings and go up till its present use. Throughout the time period, AI shifts from being symbolic reasoning systems, popularly referred to as GOFAI, to more adaptive machine learning approaches and then to the deep learning models, which are currently dominating AI. There will also be significant figures, technologies, and challenges mentioned

throughout the development of AI, and both the successes it has enjoyed so far as well as the ethical concerns that continue to be brought up will be discussed.

Early Foundations and Theoretical Beginnings. The 1940s and 1950s is when computer scientists such as Alan Turing and John von Neumann began to consider if it was even possible for machines to think like humans. It was Turing who, in his seminal paper "Computing Machinery and Intelligence" in 1950, introduced the concept of the Turing Test, which had a way to determine whether or not a machine would be capable of exhibiting intelligent behavior indistinguishable from a human. Computer scientists like Alan Turing and John von Neumann started off the path to replicating human intelligence by first attempting to define intelligence in a machine, and this set up the next couple decades of research trying to replicate the intelligence of humans.

Symbolic AI and Expert Systems. Main areas of research during the 1960s and 1970s were focused more on symbolic AI, or so-called "Good Old-Fashioned Artificial Intelligence," abbreviated as GOFAI, like in the introduction of Section 2. Marvin Minsky and John McCarthy are among those scientists who built systems capable of solving particular problems through symbolic reasoning and application of logic. This is the age in which expert systems have been developed with given rules and knowledge bases to simulate decision-making in a particular domain. A system such as DENDRAL (for chemical analysis) or MYCIN (for medical diagnosis) really demonstrated the potential for a greater degree of AI in specialized activities. Yet these kinds of systems could not scale and adapt to deal with the complexity and variability characteristic of real-world settings [30].

The AI Winter and the Shift to Machine Learning. The early successes gave way to the realization of the limitations of symbolic AI, followed by periods of reduced funding and interest that were termed the "AI Winters." Beginning in the 1980s, researchers moved away from rule-based systems and toward approaches more firmly set within conceptions of human learning. This marked the beginning of a new era: Machine Learning (ML). These new ML computer programs were able to learn and get better on their own, rather than needing to rely on pre-existing data. Researchers like Geoffrey Hinton and Yann LeCun started developing neural networks, which could learn from examples and were able to spot patterns and make predictions all on their own. This, along with faster computers and lots of data, helped bring about today’s advanced AI.

Deep Learning and Modern AI. There was a re-awakening in the field of AI during the 2000s and 2010s: deep learning. These models are able to leverage multilayer neural networks in successfully completing tasks such as image recognition, natural language processing, and game playing. Success for AlexNet in the 2012 ImageNet competition marked the CNN renaissance in

computer vision, while models such as Google's AlphaGo demonstrated deep reinforcement learning applied to complex strategic games. Indeed, the perfect storm of big data, an increase in computational power such as GPUs and TPUs, and sophisticated algorithms have launched AI applications from the realm of theoretical research into mainstream applications like autonomous vehicles and natural language understanding [1-3].

While great progress in this domain has been achieved, many challenges have also been encountered. Some of these issues regard problems with machine learning models and bias, explanatory transparency or, rather the so-called "black box" problem, and ethical concerns regarding autonomous systems. The AI systems, built upon prejudiced data, may create stereotypes or discriminate against one or more groups by propagating partiality. Also, sensitive areas such as health, law enforcement, and financial services call for careful considerations of ethical guidelines and regulatory frameworks concerning AI deployment. The overcoming of these challenges is especially significant in view of further development and implementation of AI technologies in a responsible way [31].

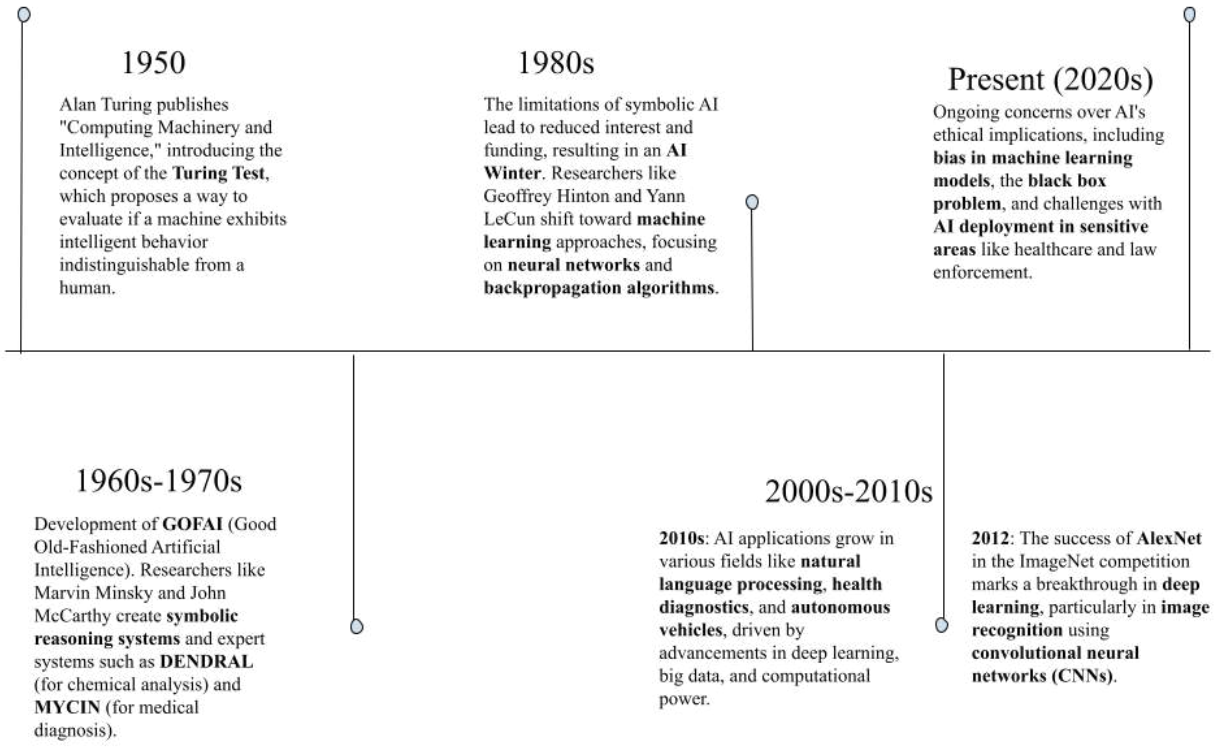


Fig 1: AI History Timeline

Section 3. (Defining) The human intelligence

This section is going to investigate the interface of human intelligence and rational architecture; how the framing of intelligence on both biological and theoretical levels is done. It will explain in detail how such a concept of the rational architecture of the human mind explains intelligence as an emergent property through a philosopher named John Pollock: the interaction of beliefs, desires, and intentions. Section 3 will look at how intelligence has been defined by introducing the concept of degrees of personhood, how both humans and AI systems can be intelligent, and how practical and theoretical reasoning forms one aspect of this complex process.

The Evolution of Human Intelligence in Science. The scientific investigation into human intelligence started in the early 20th century when psychologists, including Charles Spearman, began examining the concept of general intelligence, or g, that postulated that a single, underlying cognitive ability influences all mental tasks. This was the beginning of psychometric theories, where there was quantification of intelligence through mental performance tests, such as IQ tests. The concept of intelligence in due course has changed, or rather it accommodated its complexity, and further new models emerged challenging the traditional single view of entity conceptualization of intelligence.

Today, it is viewed as a composite of various abilities that one possesses to navigate through various environments. The theory on multiple intelligences by Howard Gardner expanded the scope to include the identification of separate intelligences-musical, interpersonal, and kinesthetic intelligences among others-which could not be defined by regular IQ tests. This disposition of comprehension projects a broad outlook toward adaptability in the brain and diversity in intelligence; hence, scientists continuously update what has been thought to be the limit of human intelligence [32].

Frameworks for Measuring Human Intelligence. The psychometric approach to intelligence testing remained the dominant one for most of the 20th century, with tests like Stanford-Binet and Wechsler Adult Intelligence Scale or WAIS which had reduced intelligence to a single numerical score. These types of tests measure verbal reasoning, memory, problem-solving abilities, and other cognitive skills; these critics have claimed that narrow logical-mathematical abilities are the focus of such tests.

In contrast, Gardner insists on multiple intelligences that do need alternative modes of assessment in capturing creativity, social skills, and other non-traditional forms of intelligence. Robert Sternberg's triarchic theory provides an expanded framework; it focuses on three varieties of intelligence - analytical, creative, and practical - all of which serve in ways that individuals solve problems in the world around them. Indeed, Sternberg himself did work that impacts modern approaches that focus on real-world application of intelligence. These widen the scope of testing to include such things as non-academic skills [33].

Other theories include the concept of emotional intelligence or EQ, brought about by Daniel Goleman, which is one's capability to handle or control emotions and deal with social

settings. It is evidence that intelligence influences larger aspects of life other than intellectual ones. For that matter, emotional intelligence is often assessed through self-report measures or behavioral assessments that appraise how individuals recognize and regulate their feelings; hence, it became vital in areas involving leadership and education [4-8].

The Role of Genetics and Environment in Human Intelligence. One of the driving forces in the field of researching human cognition is the long-standing debate of how much of human intelligence is determined by genetics and how much is determined by the environment that surrounds the human. Results from twin studies and other twin research have pointed toward DNA as the basis for the capability of an individual in intelligence, hence showing that there is genetic influence on intelligence. These findings do not, however, also rule out the critical role played by environmental influences such as education, culture, and socio-economic background that go a long way in building up intelligence [34].

Evidence from research documents access to quality education as one of the important elements driving improvement in cognitive capability over time. To this effect, psychologists such as Raymond Cattell have described fluid and crystallized intelligence, where fluid intelligence describes one's ability to solve new problems and adapt to novel situations, while crystallized intelligence reflects knowledge acquired from life experiences. While fluid intelligence tends to decline with age, crystallized intelligence is raised showing the interaction of biological and experiential factors across the course of development in intelligence [35].

The Human Brain and Rational Architecture. The human brain is the physical basis of intelligence, but it is more complex than the interconnections among its neurons. Theoretically, the mind is framed on "rational architecture" according to John Pollock's "How to Build a Person," with a conceptual model for mapping functional organization [5]. This author believes that what sets the human brain apart is a form of intelligence defined by theoretical and practical reasoning. This much-needed reasonable architecture is the dynamic interaction of states such as beliefs, desires, and intentions whose interaction will yield in rational behavior. According to Pollock, rather than thinking of intelligence as some latent attribute, it is an emergent property that results from interactions in this architecture [5]. The capacity of the brain for abstract processing, simulating outcomes, and engaging in recursive thinking enables it to go beyond a mere stimulus-response mechanism. This reflects a degree of sophistication that is difficult to emulate with the use of artificial systems.

Rethinking Traditional Definitions of Intelligence. Traditional intelligence along with the accompanying IQ tests have come under such a barrage of attacks because the limits on understanding in logical reasoning and linguistic ability were too narrow. According to Pollock, these tests do not really catch the actual conceptualization of intelligence as they tend to disregard changes to be enforced on the rational architecture [5]. In this view, intelligence is not to be thought of as an attribute nor as a list of potentials for the performance of cognitive acts,

but it is rather a symptom of how the internal states of an agent reflect a wider pattern of rational behavior. This calls into question the long-standing tradition according to which intelligence is a native property bound to biological substrates. Instead, he says that any system, either biological or artificial, that possesses a sufficiently complex rational architecture is capable of assuring that behaviors arising which we would regard as intelligent. This redefinition makes it more plausible to consider the thought that non-human entities, such as sophisticated AI systems, are possible bearers of intelligence in their doing the functional organization which is attributed to personhood [24].

Degrees of Personhood: The Continuum of Rationality. Pollock introduces the idea of degrees of personhood: that things may be more or less persons depending on their different rational architectures. The idea is that intelligence and personhood are not all-or-nothing affairs but come in shades of gray [5]. In such a case, one animal would have the degree of rationality whereby it can use tools or show empathy, though it would not have the degree of rationality which a human does. Even though they are limited in scope, these systems might eventually be thought of as "persons" if they can reach a level of reasoning and self-awareness similar to humans [24]. AI doesn’t have to imitate human intelligence exactly but just needs to act rationally—making logical choices and interacting with its environment thoughtfully—to be considered a "person."

Practical and Theoretical Reasoning. One of the key features of Pollock's doxastic architecture is the interplay between practical and theoretical reasoning. Theoretical reasoning is concerned with the formation and evaluation of beliefs about the world. By contrast, practical reasoning serves to determine intentions and decisions in the light of wants and goals. In human thought, these two types of reasoning are strongly intertwined, since practical considerations often guide theoretical inquiry and vice versa. For an entity to be human-like intelligent, it has to have this bidirectional reasoning ability. Pollock argues that the mere capability of such introspection or self-evaluation for such an entity qualifies it as a person because such an entity would therefore be able to reflect on its states and modify its behavior in response [5]. This capacity for reflection then transcends the simple processing of information but, rather, encompasses a higher-order understanding of an individual's cognitive processes, thereby facilitating the development of concepts such as self-awareness and intentionality.

Measuring Intelligence Through the Lens of Rational Architecture. In contrast, while modern assessment models of human intelligence go beyond the narrow boundaries of the traditional IQ test and encompass a more general cognitive aptitude together with emotional intelligence, the theory by Pollock attempts to say that any adequately informed theory must investigate this underlying rational architecture in which intelligent behavior is grounded [5]. This approach would therefore involve analyzing how agents-be they human, animal, or AI system-build up their internal states and deploy reasoning processes in the course of interacting with the external world. In this vein, for instance, neuroimaging studies indeed yield a wealth of

information on the neural correlates of rational architecture and indicate how different regions of the brain are involved in forming beliefs, desires, and intentions. In the field of artificial intelligence, the attempts to mimic this architecture through algorithms and neural networks create a quirky opportunity for testing our theories about intelligence. As we rebuild these systems, implementing the functional structure of the human mind, we are able to explore the edges of what it is to be intelligent and perhaps what it is to be a person.

Section 4. Differences between Human and Machine Intelligence

In this section, I aim to discuss some rudimentary differences between human and artificial intelligence. Because AI is growing into a greater presence in our lives, it therefore becomes incumbent upon us to realize the strengths and weaknesses it has in relation to human cognitive functioning. The section focuses on underlying cognitive flexibility, creativity, emotional intelligence, and experiential learning in order to point out those characteristics which truly distinguish man from machine.

Cognitive Flexibility vs. Specialized Learning. But probably one of the most fundamental differences between human and machine intelligence is the flexibility of human cognition. Humans put themselves into novel situations, apply knowledge across contexts, and solve problems in innovative and creative ways. This cognitive flexibility is arguably due to the fact that brains use transfer learning-the process of taking knowledge learned in one context and applying it to different and often unrelated situations [23].

Unlike humans, AI models typically need to retrain on new datasets to adapt to even slightly different tasks. Whereas most AI systems excel in certain tasks they are usually trained for, this is accomplished with large datasets and predefined objectives, making good predictions or decisions [9]. An example is DeepMind's AlphaGo, which was trained to play the game of Go by processing massive volumes of previously played games. But with all the dominance it showed in Go, AlphaGo would completely fail if it were to carry out a completely different activity, say, understanding language or giving a medical diagnosis. It is precisely this specialized nature of AI that limits its adaptability, while humans can function across an astoundingly wide array of environments and challenges [36].

Creativity and Innovation. The greatest artists, scientists, and innovators disclose a peculiarly human gift: the ability to envision that which has never been and then bring it into being. This fact has been developed in research conducted showing that human creativity can be extremely linked to complex neural activities, which involve divergent thinking in that there is a capability to result in various solutions for one particular problem.

This could be proven through scientific discovery: the human power of creative thinking leads to the development of theories, like the theory of relativity or the Internet. The development needed not just logical thinking but also the presence of human conceptual thinking

beyond reality. In direct opposition, AI does not have this sort of intrinsic drive and consciousness to make those kinds of creative leaps.

Whereas AI systems have indeed come up with creative content-say, art and music-the outcomes have always tended to reflect some sort of patterning from already-occupied data. Major AI systems like OpenAI's GPT-3, which isn't the latest model, generate near-human-like text based on large existing language datasets but do not really "create" in the human sense, since the ideas generated are seldom original. In fact, these models rely on pattern recognition and statistical correlations rather than insight; that is, they lack the ability for abstract thinking or the generation of meaning outstripping data-driven output. The AI-generated artwork done through systems like DeepArt or DALL·E is impressive, but again, the creativity behind the art still really emanates from the training data provided to the AI-works of art themselves. That is quite different from human art that deals with emotional expression or some kind of philosophical exploration whereby meaning and motivation are central to creative effort [37].

Emotional Intelligence and Social Interaction. Another essential difference between man and machine is the conception of emotional intelligence: the capability to recognize and understand one's feelings and others, and to manage or handle them in an appropriate manner. Human emotional intelligence is a building block for interaction in complex social environments; hence, maintaining relationships or making empathetic decisions. According to Daniel Goleman, one of the popularizers of the concept of emotional intelligence, this trait is characterized by self-awareness, self-regulation, motivation, empathy, and social skills-all combining in a person's ability to connect with others and understand emotional cues. Humans easily combine rational and emotional thoughts into a decision-making process that besides considering logical consequences also reflects on the emotional impact such a decision would have on others.

Whereas an AI system, like ChatGPT or IBM's Watson, can mimic human-like interactions and respond to questions in a manner that would arguably appear emotionally intelligent, these systems have no sense of emotional nuances. Rather, they make use of pre-programmed rules and statistical models to produce appropriate responses, lacking consciousness or depth of feelings with which to empathize or make decisions based on understanding. Systems using AI for customer service may immediately respond to an angry customer with an apology or a solution, but it all happens without any processing of the emotional understanding. AI systems, such as affective computing, try to measure and respond to human emotions based on facial expressions, tone of voice, or body language, though this kind of technology is still in its infancy and not able to function without errors [38].

Learning from Experience vs. Data-Driven Learning. While structured data teaches nothing, man learns from life experiences, intuition, and observation. In experiential learning, human beings may make informed decisions owing to the integration of formal knowledge with life experiences. Basically, human intelligence gives them the power of learning from failures and successes, refining the approaches, and adapting the strategies in real time.

On the contrary, AI relies heavily on big datasets, not to speak of well-defined rules for learning. Conventionally, any machine learning algorithm requires thousands, if not millions, to train the model in order for it to make a reasonably correct prediction. For example, AI models working for autonomous driving, such as Tesla's Full Self-Driving (FSD) system, have done extensive training with huge amounts of road data in order to understand and learn how to handle different situations while driving. But despite that fact, even with massive data, these systems fail to handle some unexpected or rarely occurring scenarios such as handling human pedestrians acting unpredictably, which intuitively may be handled by a human driver.

While on the other hand, AI systems are unable to generalize knowledge from one domain to another without explicit retraining and lack the intuitive sense developed in humans through real-world experiences. This contrast brings out the weakness of the AI in dynamic and unpredictable environments [10-14].

Section 5. Application areas

We can see how influential AI has been in the past couple of decades when we look at AI being integrated into so many different internet applications, everything from autonomous driving technologies to even the human body through neural interfaces. Especially within the field of internet technologies, AI has become indispensable in terms of managing and analyzing the tons of data on the internet. In this section, the application areas of AI will be discussed.

Internet Technologies. Artificial Intelligence has been incorporated within search engines to provide the most accurate and appropriate responses that are highly beneficial when it comes to controlling and assessing substantial resources indirectly being created online. Machine learning-based AI algorithms such as Google RankBrain, process and interpret user searches to deliver improved natural search results based on advanced understanding of semantic information query data. It has also changed digital marketing overall, making more relevant and bespoke content an even bigger part of the picture, as discussed in research on AI's role in enhancing consumer experiences online [39].

AI Chatbots. Chatbots and virtual assistants such as ChatGPT show that AI can generate human-like text responses, and also show emotions through those texts. They’re also being massively used in customer service, virtual assistance, and even healthcare, where the AI agent will give real-time information and support, as a placeholder while the customer is waiting for a human agent. These chatbots show how AI can enhance human capability and improve access to essential services in the future.

Application of AI to the Human Body. Beyond Internet Technologies and virtual chatbots, AI is increasingly being applied to the human body, especially through neural interfaces. The direct hook-up of AI systems with the human brain through devices called Brain-Computer Interfaces takes the lead in this area of innovation. Such an interface can potentially revolutionize medical

treatments in general and, more importantly, in a situation where lost sensory or motor capabilities need to be dealt with. AI-powered prosthetics, for instance, can mimic natural limb movements by responding to brain signals, which can benefit the 22.3 million amputees globally, of which a fraction are upper body amputees. AI-driven BCIs will help advance the technology needed for treating neurological disorders and enhancing human cognitive abilities [40].

Finance Field. Algorithmic trading, personal finance, and credit underwriting are current flashpoints of AI in financial services. More specifically, AI-powered trading algorithms can crawl market information to execute trades with speed and effectiveness that human traders cannot match, identifying arbitrage situations in milliseconds. Beyond trading, this AI-driven robo-advisor starts a revolution in personal finance: It analyzes the individual's spending habits, risk tolerance, and long-term objectives while providing customized investment strategies, which evolve with real-time market conditions. In underwriting credit, AI models go deeper than the traditional credit score, diving into sources of unconventional data such as social media activity and online behavioral patterns to paint a very clear picture of creditworthiness, particularly in cases where a person's credit history may be limited. This will open up not only more channels of credit but also reduce the number of defaulters by spotting subtle signals of financial discipline. This is how AI ensures that financial institutions bring increased accuracy into risk assessment and provide very personalized services, raising the bar on efficiency and client engagement.

Also, in digital marketing, AI is concerned with user retention and lead conversion. It can take a user toward the goals of the business by using intuitive AI chatbots, intelligent email marketing, and interactive web design among other numerous digital marketing services. Several factors determine the impact of AI on digital marketing. Machine Learning, a subset of AI, refers to those computer programs that access and use data in making independent learning. It collects information from social media accounts, menus, online reviews, and websites. AI uses the compiled information to create content relevant to the audience and deliver it effectively. AI software allows in-depth online analysis of restaurants and their customers. If companies implement AI into their marketing strategy, they can more effectively use the available data and reach potential customers with more striking advertisements at the best moments [41].

Education. Education is one of the application areas of AI where AI-powered tools are able to thrive because these tools can dramatically impact how students learn and how educators teach because of the personalized experiences that these tools can offer. These systems use data to adapt to the individual needs of students through customized lessons and feedback targeted at the areas in which the students struggle the most. For instance, online learning platforms such as Coursera and Duolingo have begun to use AI, and what these platforms do is adjust the speed and difficulty of the lessons based on the student performance on previous material in order to maximize efficiency in learning. AI-powered tutoring systems are another example of how AI can be integrated into education because they also work as virtual teaching assistants as they can

answer questions that students have through a chatbot, grade assignments automatically, and even suggest extra resources based on every student's individual needs.

Section 6. Challenges and Limitations

This section develops from the basic capability of AI and potential ethical considerations to practical challenges and limitations. With AI systems being increasingly influential in major decisions in the fields of finance, hiring, and justice, the issues of bias and fairness throughout the different systems are starting to question the reliability of AI for different tasks. It will also review the complexity around the definition of fairness in AI and some of the main ethical trade-offs that face developers. This section naturally leads from earlier sections where the theoretical underpinning of AI was discussed and some ethical implications put into real-world applications, showing with great need for regulatory oversight [15-16].

The rising application of AI in critical decision making for lending, hiring, and criminal justice has brought into the limelight the problem of bias that is rapidly booming in machine learning systems. Skewed data, from which most of these algorithms derive their operation rules, sometimes produces side effects that may affect demographic groups disproportionately or even lead to discriminatory outcomes. Whereas there have been various advances in machine learning, and tools such as the introduction of IBM's AI Fairness 360, the challenge is partly that the definition and realization of fairness are complex. There are over 21 different fairness metrics a developer or policymaker may choose, each with different results, hence making it difficult to know which one would best apply to each use case. This inconsistency might cause a serious risk when these algorithms are deployed to sensitive domains with great influence on people's lives [42].

The tension lies fundamentally between group fairness, which relies on the concept of equal outcomes for diverse demographic groups, and individual fairness, which tries to treat similar individuals similarly. These objectives conflict with each other. It is tough to satisfy all metrics of fairness simultaneously, hence giving way to the challenge of competing fairness definitions. For example, an AI system built to predict recidivism in the criminal justice system would appear as following group metrics but could implicitly penalize people who fall outside of those assumptions that go into the model. But the presence of such a multitude of approaches-because there is no clear guidance around fairness-introduces severe ethical considerations into areas where AI solutions are increasingly being used to make decisions affecting social and economic opportunities [43].

To that end, AI Fairness 360 will be instrumental in embedding bias detection and correction at different levels of the AI development pipeline. These range from pre-processing techniques, which clean the data of bias, to in-processing methods that actually change the learning process itself, to post-processing strategies that correct biased outcomes. Developers can reduce the chance of producing models that perpetuate or magnify existing biases by applying these techniques early on in an AI development process. But even with these tools, developers

face tough trade-offs between fairness and model accuracy; in some instances, reducing the bias comes with performance costs.

One other proposed solution involves baking the constraints of fairness into model training, where AI systems optimize not just for accuracy but for fairness, too. It can mean that accuracy is balanced across different demographics or that protected attributes like race or gender do not unduly influence predictions. AI developers can make sure their models pass the ethical bar by implementing constraints for fairness in their models. Inclusion of such features, however, calls for continued auditing and retraining once the models go live to capture potential biases.

Concerns Regarding the Integration of AI into Human Society. One of the major concerns with integrating AI into human life is that people will lose decision-making autonomy. As AI begins to involve themselves even more so in healthcare, financial institutions, and law enforcement, some might begin to rely too heavily on machine-made decisions. For instance, AI systems such as IBM Watson being used to help doctors in suggesting treatment plans for cancer patients. While much of these systems can indeed provide insightful and data-driven suggestions, over-reliance on AI in these critical areas might result in less and less human judgment being applied, especially in areas where nuanced ethical considerations are paramount. A study published in The Lancet found that while the use of AI-assisted diagnosis increases precision, it also runs the risk of supplanting the critical thinking skills of doctors in instances where moral or complex human elements are at play. Further, in criminal justice, AI systems like COMPAS, used in the estimation of recidivism, have received criticism since such systems may lead to unjust sentences against either racial or socio-economic groupings. These systems depend on data from history that may be riddled with inequity and hence raises ethical questions about fairness and to what extent humans can trust AI in making consequential decisions.

With increasing autonomy invested in AI systems, especially in more complex decision-making tasks such as driving and medical diagnosis, one of the big questions is: who is responsible in case something goes wrong? Thinking of accidents involving autonomous cars - for example, Tesla's Autopilot system - there have been fatal ones where it is not clear whether the fault lies with the driver or with the AI system. One highly publicized incident involved a Tesla operating on Autopilot rear-ending a stopped firetruck. The accident has illustrated some of the complicated questions about blame raised when AI systems and humans divide authority over decisions. As these AI systems continue to move toward more autonomous functioning, clear legal structures and lines of accountability will be required to define liability in instances where an AI-powered system goes wrong. AI systems like DeepMind's health app have also faced scrutiny in the medical field due to data privacy concerns, with implications of disastrous consequences for patients should any medical advice go wrong. That is the challenge in terms of a regulatory environment which ensures both safety for users and responsibility for developers while sustaining benefits to enhanced decision-making capacity given by AI.

Long-term psychological implications, too, are a cause for grave concern with the integration of AI in human life. While AI appears to be performing almost all functions, from personal assistants like Siri and Alexa to AI-powered educational platforms, there is a risk that people will begin relying too much on machines even in decision-making, which ultimately weakens one's problem-solving and critical thinking skills. Meanwhile, with the continuing development of AI technologies, including BCI, ethics started to raise concerns about changing human cognition. However, in the near future, BCIs will offer humans a more direct interface to the AI computing systems such that the former will be able to enhance human cognitive capabilities or even restore those which have been lost due to accidents or diseases, for example, memory and motor capabilities. This development is poised to bring some ethical questions about what constitutes human identity and raise profound moral dilemmas regarding enhancement in human capabilities beyond what nature has endowed. This raises the question of just what the addition of artificial intelligence to human thought will do to tradition rich philosophical conceptions of human nature, not to mention the possibility of unfair advantages or uneven playing fields [44-48].

As discussed in earlier sections, integrating AI into human life brings complex philosophical questions about fairness, accountability, and identity that all require careful consideration. This section’s analysis bridges those theoretical foundations with the practical impacts of AI-human interaction. It serves to provide further information for how society can responsibly coexist with AI as it continues to grow.

Section 7. Regulations and Policies

Of all, the General Data Protection Regulation of the European Union has been one of the broadest frameworks in meting out issues principally related to AI. The GDPR provides that a person shall not be subject to a decision based solely on automated processing unless explicit consent is given. This directly addresses AI-driven decision-making systems with the purpose of assuring individuals that they are in control and informed about how critical decisions regarding them are made by AI systems. It addresses transparency by discussing that organizations must provide explanations of AI decisions when personal data is involved, therefore giving them a clear legal accountability framework.

On the other side, with regard to the United States, there have been explicit regulations put in place on the use of AI in hiring under the Artificial Intelligence Video Interview Act of the State of Illinois. Specifically, employers must inform candidates that AI will be used to screen video interviews, explain the workings of the system, and obtain the candidate's consent. It also restricts data sharing and requires that video recordings should be destroyed upon the demand of the candidate. This regulation indeed reflects the ethical need for transparency and consent in these systems, the subjective features of evaluation being features like the expression of face or tone of voice. Meeting concerns on privacy and prejudice in AI hiring, this ensures fairness to candidates.

All these policies put together will contribute to the management of those ethical challenges thrown up by AI, especially on issues of bias, transparency, and human oversight. While these are regulations with a very strong founding, they can be further refined by introducing comprehensive auditing standards into AI systems. Detection of bias can, for instance, be extended into real-time monitoring of AI decision-making processes against which instant corrections are made for any potential biases. Beyond this, the regulation could spell out in greater detail what explainability might entail, such as requiring not just transparency in the underlying decisions themselves but also with actionable feedback given to candidates as ways of improving outcomes. Finally, accountability would be increased by emphasizing more the importance of independent auditing of AI algorithms, including independent review of whether AI systems meet ethical standards without discriminatory practices [49].

Section 8. Conclusion

Given the pace at which AI is influencing the most important sections of society, understanding and addressing the challenges arising out of it becomes quintessential. Major areas in which AI integration causes a few ethical and practical problems have been discussed with regard to bias, fairness, and erosion of human autonomy throughout this paper. However, regarding this matter, while we think that technological advances such as real-time auditing for bias and fairness constraints do provide some scope for improvement, they also hint at the need for continuous monitoring and improvement. It will be an informed regulatory policy that plays an important role in making AI systems maintained to be fair, transparent, and accountable in their ongoing evolution. It will be in integration with technologists and policymakers that collaboration will assure a responsible integration of AI into the future to protect human interests and maintain ethical standards in that future.

Works Cited

https://ellow.io/components-of-ai/
https://www.researchgate.net/figure/AI-Platform-Architecture-21_fig7_332555792
https://www.restack.io/p/ai-frameworks-answer-ai-architecture-diagram-cat-ai
https://yuilihistory.com/cms/files/original/91401db58911aeb5227ff0fbed4708e9.pdf
https://www.jstor.org/stable/2214201
https://www.sciencedirect.com/science/article/abs/pii/001002778390032X?via%3Dihub
https://cococsci.princeton.edu/papers/griffithsunderstanding.pdf
https://www.cell.com/neuron/fulltext/S0896-6273(12)00584-3?_returnURL=http://linkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0896627312005843%3Fshowall%3Dtrue&cc=y=
https://www.sciencedirect.com/science/article/pii/S0079612319300615
http://www.cs.yale.edu/homes/dvm/papers/conscioushb.pdf
https://www.pnas.org/doi/abs/10.1073/pnas.1905334117
https://orbit.dtu.dk/en/publications/human-vs-machine-intelligence-how-they-differ-and-what-this-impli
https://backend.orbit.dtu.dk/ws/portalfiles/portal/201351774/bolander2019human.pdf
https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2021.622364/full
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8830968/
https://arxiv.org/abs/1810.01943
https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/
https://www.businesswire.com/news/home/20240108891905/en/Butterfly-Network-Announces-FDA-Clearance-of-its-Next-Generation-Handheld-Ultrasound-System-Butterfly-iQ3
https://www.sciencedirect.com/science/article/pii/S2949866X24000285
https://technode.global/2024/05/17/estonias-education-minister-kristina-kallas-on-the-challenges-and-opportunities-of-ai-in-learning-and-empowerment-qa/
https://www.npr.org/2024/06/07/nx-s1-4992007/how-ai-tools-are-being-used-in-classrooms
https://oecd.ai/en/wonk/evolving-with-innovation-the-2024-oecd-ai-principles-update
https://www.sciencedirect.com/science/article/abs/pii/S0079612307680108
https://www.researchgate.net/profile/Eric-Dietrich/publication/314732360_Artificial_Intelligence_Philosophy_of/links/5cc0a6fe4585156cd7afa317/Artificial-Intelligence-Philosophy-of.pdf
https://www.cp.eng.chula.ac.th/~prabhas/teaching/cbs-it-seminar/2012/aiphil-mccarthy.pdf
https://link.springer.com/article/10.1007/s13347-021-00460-9
https://link.springer.com/article/10.1007/s12559-018-9619-0
https://www.researchgate.net/profile/Christoph-Durt/publication/375006419_Artificial_Intelligence_and_Its_Integration_into_the_Human_Lifeworld/links/653ba74af7d021785f15fb01/Artificial-Intelligence-and-Its-Integration-into-the-Human-Lifeworld.pdf
https://link.springer.com/article/10.1186/s41235-023-00499-6
https://doi.org/10.1016/0016-3287(77)90097-0

https://doi.org/10.1007/978-3-658-40004-0_2
https://www.proquest.com/openview/43a4ad0419a39e3c1ea249edd0237745/1?pq-origsite=gscholar&cbl=2026366&diss=y
https://doi.org/10.1016/0273-2297(85)90017-6
https://www.nature.com/articles/mp2014105
https://www.proquest.com/openview/9464b67e34a2cb16135eb1ad70576e9b/1?pq-origsite=gscholar&cbl=18750
https://www.science.org/doi/abs/10.1126/science.aar6404
https://digitalcollection.zhaw.ch/items/d2ac28d6-3f4a-4518-af88-b0a9511f4876
https://www.tandfonline.com/doi/abs/10.1207/s15326985ep4104_4
https://doi.org/10.3390/app13148114
https://doi.org/10.1016/B978-0-444-52901-5.00006-X
https://ieeexplore.ieee.org/document/10197415
https://dl.acm.org/doi/abs/10.5555/3648699.3649011
https://link.springer.com/article/10.1007/s10940-022-09545-w
https://heinonline.org/HOL/LandingPage?handle=hein.journals/jtelhtel17&div=9&id=&page=
https://www.researchgate.net/profile/Missy-Cummings/publication/372051108_Identifying_AI_Hazards_and_Responsibility_Gaps/links/64a30a848de7ed28ba7201e1/Identifying-AI-Hazards-and-Responsibility-Gaps.pdf
https://dpl6hyzg28thp.cloudfront.net/media/If_AI_becomes_conscious_heres_how_researchers_will_know.pdf
https://www.nature.com/articles/d41586-023-02684-5
https://arxiv.org/pdf/2308.08708
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3386914

Challenges and Prospects of Student Guidance and Counseling Policies in Taiwan

By Cun-Qian Huang and Ling-Yu Wang

Abstract

This paper investigates the pressing issue of student mental health in Taiwanese junior and senior high schools, focusing on the challenges facing the school counseling system. Drawing upon recent tragic incidents and survey data, the paper highlights the high prevalence of stress and depression among adolescents, particularly among senior high school students. Despite these alarming figures, a significant proportion of students do not seek support from school counselors, indicating a systemic failure to effectively address their needs. This paper analyzes the root causes of this failure, including inadequate staffing, excessive workload for counselors, high turnover rates among professionals, the passive role of counseling offices within the school system, and a lack of clear regulations governing counseling responsibilities. To address these challenges, the paper proposes several key recommendations: (1) formalize the division of responsibilities in counseling work through clear regulations, (2) clearly define the counseling office's role in case handling procedures, (3) enhance collaboration between school counselors and external clinics, and (4) strengthen counseling-related training in teacher education programs. By implementing these recommendations, this paper suggests, Taiwan can create a more robust and effective school counseling system that better supports the mental health and well-being of its students.

Keywords

School Counselling, Mental Health, Depression, Student Wellbeing, Bullying

Statements and Declarations

The authors contributed equally to this paper.

Introduction

In February 2023, a tragic incident occurred in Taiwan when a student from Taichung Municipal Feng Yuan Senior High School died by suicide at home. Since the second year of high school, this student had been subjected to relentless bullying and harassment by the dean of student affairs and other school staff. These officials repeatedly accused the student of theft, vaping, and even drug use, despite a lack of evidence. The school subjected the student to frequent and intimidating interrogations, conducted illegal body searches, and coerced false confessions for actions the student did not commit. School officials even publicly humiliated the student, labeling them as “social trash” and “garbage” in front of their classmates. In addition to these verbal abuses, the school imposed numerous unjustified penalties, resulting in a mounting accumulation of warnings, minor demerits, major demerits, and school absences. The dean of student affairs eventually pressured the student to withdraw from school. After enduring several months of this unrelenting bullying from the school, the student tragically died by suicide (Yen).

In December 2017, a similar tragedy occurred at Taipei Municipal Min-Quan Junior High School, where a first-year student diagnosed with Tourette's Syndrome died by suicide. This student had been involved in a campus gender incident referred to the school's Gender Equity Education Committee. However, the school failed to conduct a proper investigation and did not provide the student with a clear incident report. During committee meetings, the student was compelled to explain their actions in front of others, without any consideration for their psychological condition and the potential impact on their well-being. Furthermore, the Gender Equity Education Committee overstepped its authority by bypassing the Student Reward and Punishment Committee and directly imposing a series of harsh and potentially unlawful penalties. These included social isolation from classmates, removal from the basketball club, and forced standing for extended periods as punishment.

The student was frequently sent to the dean's office, where he was subjected to loud reprimands, harsh punishments, and forced to retake exams. Despite expressing suicidal thoughts to the counseling office, the school failed to respond adequately. The counseling sessions were brief and superficial, lacking any meaningful intervention or follow-up. Meanwhile, the school continued to impose inappropriate disciplinary measures and mistreatment, exacerbating the student's distress. Tragically, after only 98 days of enrollment, the student died by suicide (Control Yuan).

What has gone wrong with the student guidance and counseling systems in Taiwan? How have schools, institutions designed to foster holistic development and safeguard student well-being, become environments that can negatively impact their mental health, even contributing to tragic outcomes like suicide? Why are counseling systems often ineffective in addressing students' mental health needs or providing meaningful support during incidents like bullying?

Beyond extreme cases, similar issues persist in broader contexts. The online platform “Crossing” conducted interviews exploring why students avoid seeking help from school counseling offices. Some students reported that peers struggling with academic pressure sought assistance from counselors but found their situation to have deteriorated rather than improved. Consequently, these students opted to seek help from external psychiatric professionals instead of relying on school counselors. Other students noted that some counselors often oversimplified student concerns, dismissing anxieties about academic pressure with dismissive advice. This approach failed to address their genuine emotional needs, leaving them feeling neglected and misunderstood. This often compelled them to seek support from external medical resources (Liao).

Why has student counseling reached a crisis point, failing to adequately meet the needs of today's students? How can we reform and improve this crucial system? This article aims to explore these central questions.

Understanding the Mental Health of Taiwanese Adolescents Through Data

To understand the general mental health status of Taiwanese adolescents, we examine the

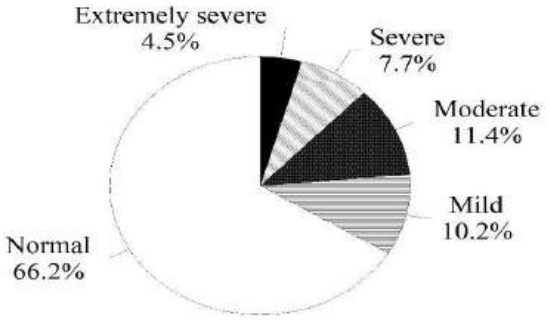
findings of the “2023 National Survey on the Mental Health of Junior and Senior High School Students” conducted by the Child Welfare League Foundation of Taiwan. This survey employed cluster proportional sampling, dividing Taiwan's 21 counties and cities into four regions: north, central, south, and east. From May 31 to June 30, 2022, the foundation surveyed junior high school students (grades 7-9, aged 12-14) and senior high school students (grades 10-12, aged 15-17) using an online questionnaire. The sample size was determined by the population ratio in each region, resulting in 1,842 valid responses. After repeated weighting adjustments, the sample composition comprised 52.2% male and 47.8% female respondents, with 48.5% being junior high school students and 51.5% being senior high school students (Child Welfare League Foundation).

The survey utilized two subscales – “Depression” and “Stress” – from the DASS-21 (Depression, Anxiety, and Stress Scale-21) to assess the levels of depression and stress among junior and senior high school students.

Stress Levels Among Junior and Senior High School Students

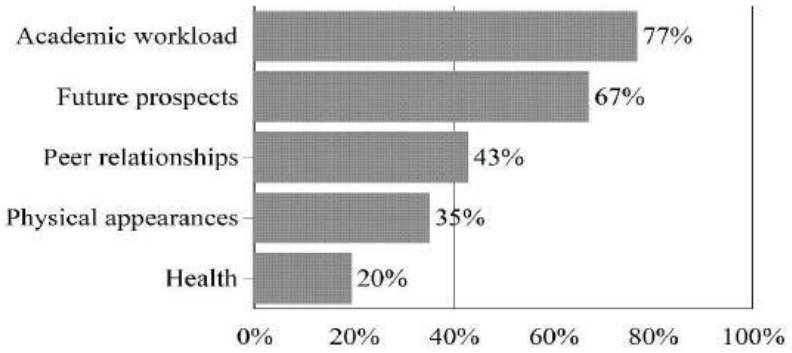
The survey utilized the “Stress” subscale of the DASS-21 Emotional Self-Assessment Scale, which included items such as “In the past week, I felt: it was hard to calm down, irritable and sensitive, restless, difficult to relax, and mentally drained daily,” among a total of seven items. The results indicated that 12.2% (4.5%+7.7%) experienced stress levels classified as severe or higher, as shown in Figure 1. To distinguish between junior and senior high school students, the percentage experiencing a severe stress level or higher is 8.2% and 16%, respectively.

Figure 1
Stress Levels Among Junior and Senior High School Students in Taiwan
2023 National Survey on the Mental Health of Junior and Senior High School Students



Regarding the sources of stress for junior and senior high school students, the survey revealed the top three causes: academic workload (76.9%), future prospects (67.3%) and peer relationships (43.0%), as illustrated in Figure 2.

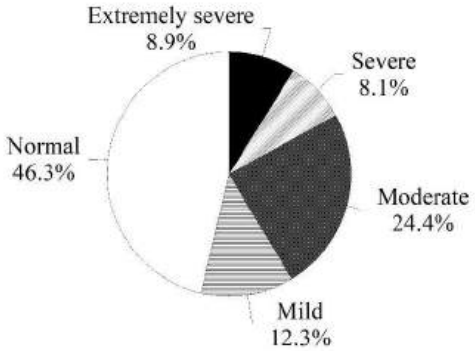
Figure 2
Sources of Stress for Junior and Senior High School Students in Taiwan
2023 National Survey on the Mental Health of Junior and Senior High School Students



Depression Levels Among Junior and Senior High School Students

The survey utilized the "Depression" subscale of the DASS-21 Emotional Self-Assessment Scale, which included items such as, "In the past week, did you feel: unable to experience positive emotions, difficulty starting tasks, a lack of things to look forward to, a sense of worthlessness, or feeling depressed and down," among a total of seven items. The results revealed that 17.0% of high school students exhibited depression levels classified as severe or higher, as illustrated Figure 3.

Figure 3
Depression Levels Among Junior and Senior High School Students in Taiwan
2023 National Survey on the Mental Health of Junior and Senior High School Students

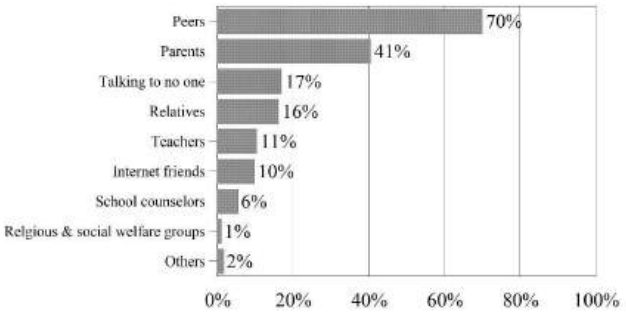


A distinction between junior and senior high school students reveals a percentage experiencing a severe stress level or higher of 12.9% and 22.9%, respectively.

Confidants for Junior and Senior High School Students

A significant proportion of junior and senior high school students experience severe levels of stress and depression. When faced with difficulties, the survey revealed that peers were the primary confidantes for students, with 70% turning to them for support. Notably, parents ranked second, but less than half of the respondents (40.6%) sought their guidance. Teachers accounted for only 10.5% of confidantes, while school counselors were cited by a mere 5.6%, as illustrated in Figure 4.

Figure 4
Confidants for Junior and Senior High School Students in Taiwan
2023 National Survey on the Mental Health of Junior and Senior High School Students



Exploring the Underlying Concerns Behind the Data

Based on the survey results, over 10% of junior and senior high school students experience severe levels of stress and depression, with the proportion among senior high school students being double that of junior high school students. While the DASS-21 scale is not a formal clinical diagnostic tool, scores reaching moderate levels or above indicate a psychological condition that warrants attention and suggests the need for professional assistance.

The survey also reveals that the mental health needs of junior and senior high school students may not be adequately addressed through proper and professional channels. Among the individuals students confide in during times of distress, only 5.6% reported seeking help from school counselors, a figure even lower than the 9.8% who turn to online acquaintances. Even general schoolteachers were approached by only 10.5% of students. Clearly, a significant proportion of students require professional mental health support, yet the school counseling system fails to effectively provide the necessary assistance.

Challenges in Student Counseling Work

Problems rarely arise from a single perspective. Just as students in need of psychological support may not receive adequate help, the student counseling system itself faces significant challenges. What obstacles hinder the effective functioning of student counseling services? Let's explore the answers below.

Inadequate Staffing in Counseling Services

According to the *Student Guidance and Counseling Act* of Taiwan and its Enforcement Rules, junior and senior high schools are mandated to allocate one "full-time guidance counselor" for schools with up to 12 classes, two counselors for schools with 13 to 24 classes, and so forth for larger institutions. However, in practice, many schools fail to meet the professional counseling staffing standards outlined in this legislation (Ministry of Education).

This shortfall likely contributes significantly to the heavy workload experienced by counselors. Firstly, counselors are often responsible for an overwhelming number of students, making it challenging to provide effective and personalized support. Secondly, in smaller schools with fewer than 12 classes, a single counselor may be assigned to serve two or even three schools simultaneously. This arrangement necessitates frequent travel between campuses, hindering the provision of consistent support to students and obstructing the development of long-term, trusting relationships.

Furthermore, schools in remote areas face unique challenges, such as inadequate transportation infrastructure, further complicating the implementation of effective counseling services (Legislative Yuan 130–131).

Excessive Workload for Counselors in Campus

According to the Ministry of Education's *Operational Manual for Professional Services Related to Special Education*, the role of counselors is described as follows: "Counselors serve all students in the school, offering support through information dissemination, class counseling, lectures, workshops, and tests to help students address various developmental issues, such as career planning, interpersonal relationships, learning methods, and family dynamics" (Wang 54). Specifically, the main responsibilities of counseling offices can be categorized into three major areas:

Three-Tiered Counseling System

Counseling responsibilities encompass a spectrum, including primary counseling (e.g., regular classroom guidance and addressing general issues) and more intensive secondary and tertiary counseling for high-risk individual cases. With the increasing diversity and volume of student issues, counselors often find themselves managing multiple cases simultaneously.

Academic Counseling

This involves assisting students with aptitude and interest assessments, conducting one-on-one interviews, organizing academic counseling seminars, preparing application materials, and introducing various admissions pathways and university programs. Balancing the demands of psychological counseling with the increasing focus on academic counseling presents a significant challenge for counselors. Overemphasizing one aspect at the expense of the other can have a detrimental impact on students' overall well-being.

Administrative Responsibilities

In addition to their core professional counseling duties, counselors are often burdened with administrative tasks such as planning enrollment activities and organizing student events. These additional responsibilities, often without extra compensation, significantly increase their workload and contribute to increased stress levels.

High Turnover Rates Among Full-time Professional Guidance Counselors

Full-time professional guidance counselors (different from the "full-time guidance counselors" mentioned above), typically required to hold qualifications as psychologists or social workers, primarily focus on student psychological counseling, crisis intervention, and family communication within the school setting. However, due to the high volume of cases on campus, counselors are often pressured to close cases quickly to accommodate new ones, frequently resulting in a lack of in-depth case handling and a decline in counseling quality (Legislative Yuan 139). Furthermore, professional guidance counselors face limited job flexibility, constrained by school regulations and fixed working hours. Their monthly salaries typically range from NT$40,000 to NT$50,000, offering limited opportunities for professional growth and advancement (Legislative Yuan 149).

In contrast, freelance psychologists or social workers enjoy greater flexibility and autonomy. They can choose the number and type of cases they take on, allowing for in-depth exploration of each case. Their income grows in proportion to their caseload, with high earners potentially exceeding NT$100,000 per month (Chiu, 2022). Importantly, freelancers have greater control over their workload and stress levels, avoiding the high case volume and pressure to close cases that often plague professional guidance counselors on campus.

Under these circumstances, professional guidance counselors face the dual challenges of high stress and limited salary growth, making it difficult to maintain high-quality counseling services. Consequently, many are drawn to the prospect of transitioning to freelance roles. This trend contributes to low retention rates and high turnover among campus counselors, undermining the stability of campus counseling services and negatively impacting student well-being.

Counseling Professionals as Passive Participants in School Systems

Within the realm of student affairs, the "Office of Student Affairs" typically assumes primary responsibility. However, its function often leans heavily towards disciplinary measures rather than a counseling-oriented approach. Conversely, the Counseling Office, designated as the primary unit responsible for counseling, frequently plays a passive role. While counselors may provide counseling services and maintain relevant records, their role in addressing student affairs is typically advisory, with limited influence over disciplinary decisions.

In essence, the effectiveness of counseling efforts is significantly diminished if the Office of Student Affairs issues inappropriate disciplinary actions. For example, following the tragic suicide of a student at Taipei Municipal Minquan Junior High School, the "Humanistic

Education Foundation" of Taiwan highlighted that despite the school's counseling staff making strong efforts to overturn the school's improper disciplinary actions, they lacked the authority to challenge the decisions made by the Office of Student Affairs (Legislative Yuan 144).

Lack of Clear Regulations on Counseling Responsibilities and Procedures

The *Student Guidance and Counseling Act* provides a broad overview of student counseling work but lacks specific details regarding responsibilities and procedural guidelines. While the *School Counseling Work Reference Manual* offers more specific guidance for various school levels, it lacks explicit definitions for the division of responsibilities and is not legally binding. This results in a lack of clear standards for counseling work across schools, hindering the establishment of a consistent framework for conducting counseling tasks and fostering interdepartmental collaboration for a comprehensive student counseling system. Without mandatory enforcement, schools may prioritize administrative considerations, often neglecting crucial counseling efforts.

Furthermore, the lack of effective communication channels between departments exacerbates these issues. For instance, when the Office of Student Affairs handles disciplinary cases, they may overlook the student's psychological state and proceed with rigid disciplinary actions, rarely referring these cases to the counseling office. In such instances, the counseling office remains uninformed, hindering the implementation of necessary counseling interventions.

How to Improve the Current Challenges?

The Regulations Should Clearly Define the Responsibilities of Counseling Work

The *Student Guidance and Counseling Act* lacks explicit provisions for the division of responsibilities in counseling work. Currently, only the non-binding *School Counseling Work Reference Manual* offers relatively clear guidelines for specific counseling tasks. To establish a more comprehensive framework for the division of responsibilities in counseling work, it is crucial to formalize the key points outlined in the current reference manual into official implementation guidelines. Moreover, the government should actively urge schools to strictly adhere to these regulations to ensure the effective delivery of counseling services in all educational institutions.

The Roles and Responsibilities of the Counseling Office Should Be Clearly Defined in Case Handling Procedures

In many school incident handling processes, relevant regulations fail to clearly define the roles and responsibilities of the counseling office. This often results in situations where professional counseling staff are excluded from, or even unaware of, critical cases. The tragic suicide of a student at Taichung Municipal Feng Yuan Senior High School serves as a poignant example. The accusations made by the Office of Student Affairs—such as theft, vaping, and drug use—fall under the purview of the *Juvenile Delinquency Prevention and Counseling Measures*.

These measures clearly mandate that the school should have provided counseling interventions.

However, due to the lack of explicit regulations outlining the counseling office's authority in such cases, the counseling office was entirely unaware of the student's situation. Consequently, the Office of Student Affairs mishandled the case, subjecting the student to unbearable psychological distress and ultimately leading to their tragic demise.

Therefore, it is imperative to clearly define the counseling office's roles and responsibilities within the regulatory framework for incident handling processes. This will ensure the full participation of professional counseling staff in case management, thereby preventing mishandling and averting similar tragedies in the future.

Collaboration Between Student Counseling Services and External Clinics

In response to the growing demand for psychological services among young people in recent years, the Ministry of Health and Welfare of Taiwan launched the "Mental Health Support Program for Young People" in 2023. This program partnered with psychological counseling clinics to provide three free counseling sessions to individuals aged 15 to 30. By the end of June 2023, the program had successfully served 29,920 individuals, achieving a remarkable 96% satisfaction rate.

Building upon the success of the initial program, the new "Mental Health Support Program for Adolescents and Adults Aged 15 to 45" has expanded its target demographic. To enhance accessibility, the number of contracted clinics has significantly increased, surpassing 500 by 2024. Moreover, the capacity of individual clinics has been expanded from 8 individuals per week to 12 (Ministry of Health and Welfare).

The program's high utilization and satisfaction rates underscore a genuine and significant demand for psychological counseling among young people and effectively demonstrate the feasibility and effectiveness of collaborating with external clinics. The central government could consider adopting a similar operational model for schools. By establishing a robust partnership between schools and external counseling institutions, this approach could serve as an effective referral system to address the critical shortage of professional counseling staff on campuses. Such a system would not only alleviate the workload of school counselors but also provide students with more comprehensive and accessible mental health support.

Enhancing Student Counseling Training in Teacher Education Programs

Examining the "Curriculum and Credits Table for Education Professional Courses in Pre-Service Education for Secondary School Teachers at National Taiwan Normal University" reveals that courses crucial for student counseling, such as "Principles and Practices of Counseling," "Studies on Adolescent Issues," and "Adolescent Psychology," are categorized as electives. This pattern is consistent across other universities, suggesting a potential gap in the counseling knowledge of general teachers.

To address this, universities should actively promote and encourage students to enroll in these courses, or even consider making them mandatory. Concurrently, schools should recognize

these counseling-related qualifications as valuable assets during teacher recruitment. This dual approach would incentivize teachers to acquire essential counseling knowledge. Equipped with such knowledge, they would be better prepared to handle student-related matters with greater confidence and competence, minimizing potential biases and misunderstandings towards students and the counseling profession.

Conclusion

The two recent student suicide cases, coupled with related survey data, underscore the urgent need to address the mental health of junior and senior high school students in Taiwan. Notably, senior high school students face more severe challenges than their junior high school counterparts. Despite this, a disconcertingly low proportion of students seek professional counseling, suggesting a significant gap in the effectiveness of school counseling services. This points to underlying structural issues within the school counseling system.

Several factors contribute to these problems. Beyond issues such as low salaries, heavy workloads, and insufficient staffing, a critical flaw lies in the systemic overshadowing of counseling professionalism by administrative power. The opinions of school counselors are frequently disregarded by administrative decision-makers, diminishing their influence. Furthermore, the lack of clear regulatory guidelines on the division of counseling responsibilities and established procedures hinders effective interdepartmental collaboration, leading to students' psychological needs being neglected within bureaucratic processes.

To address these challenges, a comprehensive reevaluation of student counseling regulations is imperative. The critical contents of the non-mandatory “Reference Manual” should be incorporated into formal legislation, explicitly defining the responsibilities and authority of school counseling offices. Moreover, it is crucial to ensure the active involvement of counseling personnel in handling major cases to prevent tragedies arising from purely administrative decisions. Exploring collaborations with external clinics and institutions can also help alleviate the workload of school counselors.

Additionally, teacher training programs must be enhanced to equip general educators with essential counseling skills, aligning with the *Student Guidance and Counseling Act*'s mandate that “school principals, teachers, and professional counselors all bear responsibility for student counseling.” Only by establishing a robust and clearly defined counseling support system can we effectively improve students' mental health and foster their holistic physical, emotional, and spiritual development.

Works Cited

嚴文廷（2023年4月24日）。專訪豐原高中生父親：一個好老師可以救學生，但我兒子在豐中沒遇到。報導者。https://www.twreporter.org/a/taichung-feng-yuan-senior-high-school-bullying (Yen, W.-T. “Interview with the father of a Feng Yuan Senior High School student: A good teacher can save a student, but my son didn't meet one at Feng Yuan.” *The Reporter*, 24 April 2023.)

監察院（2019月5月16日）。臺北市民權國中在**106**年**12**月間，發生一名妥瑞氏症學生於放學後自住家跳樓身亡事件，監察院對此調查發現，民權國中處理該生就讀期間之性平事件，程序明顯違法，且該校輔導管教該生之作法不當，監察院對民權國中提出糾正，並要求重新檢討該校相關人員疏失責任；此外，針對我國各級學校學生自殺/傷行為攀升趨勢，監察院呼籲教育部正視並有效處理。https://www.cy.gov.tw/News_Content.aspx?n=213&s=13420 (Control Yuan. *An investigation by the Control Yuan revealed that Taipei Municipal Min-Quan Junior High School had serious procedural violations in handling a gender equality case involving a student with Tourette's syndrome. The student tragically died by suicide after jumping from their home in December 2016. The Control Yuan also found that the school's counseling and disciplinary actions towards the student were inappropriate. As a result, the Control Yuan issued a corrective action to the school and demanded a thorough review of the negligence of the relevant personnel. Additionally, given the rising trend of student suicides and self-harm in schools nationwide, the Control Yuan urged the Ministry of Education to address this issue seriously and effectively*, 16 May 2019.)

廖宥甯（2023年9月4日）。求助無門的青春：為什麼學校已有輔導室，這群高中生還得「自尋生路」？。換日線。https://crossing.cw.com.tw/article/18073 (Liao, Y.-N. “The desperate youth: Why do these high school students have to seek help on their own when schools already have counseling offices?” *Crossing*, 4 Septermber 2023.)

兒童福利聯盟（2023年2月）。**2023**年臺灣國高中生心理健康調查結果。https://www.children.org.tw/publication_research/research_report/2544 (Child Welfare League Foundation. *Results of the 2023 mental health survey of junior and senior high school students in Taiwan*, February 2023.)

教育部（2024年1月4日）。教育部召開首次校園安全諮詢會 就三大議題與教師家長及學生團體代表交換意見會商。教育部全球資訊網。https://www.edu.tw/News_Content.aspx?n=9E7AC85F1954DDA8&s=DC51E69A09866A7E (Ministry of Education. *The Ministry of Education convened its first campus safety advisory meeting to discuss three major issues and exchange opinions with representatives from teachers, parents, and student groups*, Ministry of Education Global Information Network, 4 January 2024.)

立法院（2024）。立法院第11屆第1會期教育及文化委員會「學生輔導法」修法公聽會紀錄。立法院公報，113(27)，123-226。(Legislative Yuan. Minutes of the public hearing on the amendment of the Student Counseling Act at the 1st session of the 11th Legislative Yuan,

Education and Culture Committee. *Legislative Yuan Publication, 113*(27), 2024, pp. 123-226.)

王天苗（2003）。特殊教育相關專業服務作業手冊。臺北市：教育部特殊教育工作小組。(Wang, T.-M. *Handbook of Specialized Educational Services*. Taipei City: Ministry of Education, Special Education Task Force, 2003.)

邱意婷（2022年7月4日）。坐在優雅空間、聊聊天，就能賺大錢？──破除那些你對「心理師」的誤會。換日線。https://crossing.cw.com.tw/article/16432 (Chiu, Y. T. "Sitting in an elegant space, chatting, and making a fortune? - Debunking the misconceptions you have about psychologists." *Crossing*, 4 July 2022.)

衛生福利部（2024年8月1日）。青壯的心誰傾聽？心理健康支持擴大方案來了！衛福部「**15-45**歲青壯世代心理健康支持方案」**8**月**1**日上路。衛生福利部。https://www.mohw.gov.tw/cp-16-79408-1.html (Ministry of Health and Welfare. *Who listens to the hearts of young adults? expanded mental health support program is here! The Ministry of Health and Welfare's "Mental Health Support Program for Young Adults Aged 15-45" is launched on August 1st*. 1 August 2024.)

Trade Turbulence in the Eurozone: The Impact of Economic Integration and Currency Unification on Cross-Border Volatility By Saaj Shah

Abstract

In this paper, I explore how joining the European Union and adopting the EURO as the official currency affects cross-border trade volatility. This study employs trade metrics such as imports and exports as percentages of GDP and in constant 2015 US dollars, alongside recent real GDP per capita data, to identify correlations and study patterns. The findings reveal that adopting the euro and EU membership increased trade volatility, particularly for imports (% of GDP) within the eurozone. These results shed light on the complex relationship between economic integration, currency unification, and trade dynamics, offering insights into the potential challenges of enhanced economic interdependence.

Keywords

European Union; Euro currency; Trade volatility; Imports and exports; Economic integration

1 Introduction

On February 7, 1992, the Maastricht Treaty was signed in Maastricht, Netherlands. The treaty, now officially called the Treaty on European Union laid the foundations for the European Union (EU). The Maastricht Treaty was the product of years of discussions between European governments. Finally, it came into force on November 1, 1993, serving as one of the most ambitious economic and political unions in modern history to foster integration, economic growth, and political cooperation and support amongst its member states. One of the cornerstones of the Maastricht Treaty was the advent of the euro, a single currency that has been adopted by 20 of the 27 EU member states, introduced in 1999. The geographic region that encompasses the euro is currently referred to as the eurozone or the Economic and Monetary Union (EMU). The euro was designed to eliminate exchange rate fluctuations, lower transaction costs, and enhance price transparency, attempting to spur intra-European trade and strengthen the region's economic cohesion.

One of the main factors of the EU's success and the euro's success is cross-border trade as member states rely on open markets and seamless exchange policies. However, similar to all other regions, trade is inherently subject to volatility- fluctuations in the value and volume of goods and services exchanged between countries over a specific period, influenced by factors such as economic conditions, policy conditions, and market dynamics. Trade volatility is a crucial concept because it directly affects economic stability, business planning, and global market dynamics. Additionally, fluctuations in trade can disrupt regional supply chains, alter investment decisions and strategies, and pose a threat to governments attempting to maintain a stable economy.

Despite the importance of trade stability, previous research has not thoroughly examined the effects of EU membership and euro adoption on cross-border trade volatility, particularly in developed versus developing member states. This paper addresses this gap by analyzing the impact of joining the EU and adopting the euro on trade volatility, using standard deviations of trade data before and after membership. Key metrics include percentage changes in exports and imports as a share of GDP and in constant USD, analyzed about each member state's GDP. Correlation methods will also explore the relationship between GDP per capita and trade volatility, providing insights into how economic development influences changes in trade stability.

2 Literature review

The introduction of the euro in 1999 has ignited significant academic interest in the role of the currency and how it has impacted economic growth, exports, overall economic stability, and its comparison to different currencies.

In the short period after the emergence of the euro, the Eurozone lacked the integration and flexibility required in 2004 to be an optimal currency area, demonstrating a "rigidity trap" of monetary consolidation and fiscal policy that slows labor markets. (Silvia and Stephen, 2015). Past research indicates that five years after the introduction of the euro there were economic benefits such as low interest rates and low inflation; however, it did not outweigh costs and did not catalyze growth as perceived. Instead, the primary function of the euro was to politically unify the countries of the Eurozone rather than being based on a clear economic-benefit analysis. (Rich, Georg). In the decade following the advent of the euro, research shows the move to a single currency in 1999 was a success, establishing the euro as a major international currency and increasing the credibility of the European Central Bank (ECB). Although the EMU's first decade was generally seamless, the current economic downturn has exposed different vulnerabilities across member states, prompting fears about potential tensions. However, the euro's historical endurance and mutual benefits imply that a stronger union may emerge, reflecting the United States' shift toward federalism in reaction to crises (Buti, Marco, and Paul van den Noord). Furthermore, as published in December 2009, the euro had made progress in challenging the role of the dollar as a store of value. However, findings suggest that the dollar remained the superior currency because of its incumbency advantages and the issues within the coordination of the Eurozone (NORRLOF, CARLA). Additionally, studies indicated that by 2012, the euro had not produced the anticipated level of economic stability, with academics blaming the failure on the difficulties of enforcing a single currency on a diverse group of nations. Sovereign debt crises, weak banks, high unemployment, and significant trade deficits were among the negative economic effects, which together fueled instability inside the eurozone (Martin Feldstein). Past studies have also focused on the Greek debt crisis. Researchers had come to the conclusion in 2013 that intervention was necessary to manage unsustainable policies and correct past shortcomings. Ultimately, academics posited that investment in income-generating assets, as opposed to temporary loans, was essential to lifting Greece and also

other southern European countries out of recession and maintaining economic stability (DARVAS, ZSOLT). Additional research in 2013 showcased that the push for financial centralization remained contentious because of political resistance and worries about sovereignty continued to highlight the imbalance between economic and national independence (Smith, Roy C). Furthermore, researchers went beyond the scope of the effect of the euro on the Eurozone as a whole and focused on individual regions as well. Latvia, for example, after joining the Euro had an impact on price differences where notable price differences between Germany collapsed significantly, supporting the Eurozone's ultimate goal of price harmonization (CAVALLO, ALBERTO). In 2015, a study discovered that varying monetary policies and economic recoveries influenced the global use of the euro, leading to a rise in demand for debt denominated in euros and a change in foreign investment toward assets with higher yields (European Central Bank). Next, studies found in 2016 found that adopting the euro indirectly causes economic growth in Eurozone countries by sparking a process between financial growth, debt, and improved fundamentals, which non-euro members lack. However, statistics show there are risks with over-borrowing, as seen in post-2008 when high-debt countries' economic downturns due to the reversal of this growth cycle (Kalaitzoglou, Iordanis, and Beatrice Durgheu). In addition, studies in 2018 focused on the euro's effect on systemic growth. Researchers concluded that the euro had not prompted systemic growth and the growth rate of the Eurozone is comparable to those of non-EU countries such as the UK, Sweden, and Denmark and did not outperform countries like Canada, Switzerland, and Australia. Overall adopting the euro did not provide any clear economic advantage in terms of economic growth, as the Eurozone's growth rate was relatively similar to or lower than many control group economies concluded from data (Ioannatos, Petros E). Furthermore, recent research in 2021 focused on the euro's economic stability and also its resilience throughout several crises since the 1990s. The research concluded that the euro has been more resilient than previously thought, with increasing public support and enhanced economic stability following the 2008-2013 crises. However, fresh obstacles from the epidemic persist, and future adjustments to strengthen the eurozone's resilience face persistent difficulty in reaching a consensus. (Lane, Philip R). As the pandemic became a significant issue in the early 2020s, research has focused on its impact on the euro in comparison to other currencies. Academics discovered that to strengthen the euro's international significance, Europe requires a strategic growth strategy centered on post-crisis recovery, which includes significant green expenditures. It also discovered that a European COVID-19 recovery program is required to avoid a long-term economic downturn, with green bonds and investments critical to increasing the euro's appeal to international investors (Claeys, Grégory, and Guntram B. Wolff). Studies have also focused on the intentions of the digitization of the euro of the euro finding that the digital currency seeks to reconcile the disruptive impacts of digitalization on payments with the necessity for a unified European payment system, maintaining monetary uniformity while coexisting alongside private digital currencies. The success of the initiative will hinge on regulatory coherence, how digital assets are classified, and the equilibrium among

competition, innovation, and financial stability in the developing European payments landscape (Brunnermeier, Landau)

While previous research has focused on topics such as trade, economic growth, stability, its role in comparison to other countries, and its response to crisis, in my paper I am going to research if there are any long-term implications of the creation of the European Union in 1993 and the advent Euro in 1999 on cross-border trade volatility within the Eurozone and analyze if it correlates with the Gross Domestic Product (GDP) of EU member states.

This paper follows this format. The data and methodology are described in Section 3, which also includes the most recent GDP per capita statistics and the trade metrics that were utilized, such as imports and exports expressed as percentages of GDP and in constant 2015 US dollars. It describes how to find connections and examine trends in the volatility of commerce between EU and eurozone member states. The results are shown in Section 4, which focuses on imports (as a percentage of GDP) in the eurozone and describes how EU membership and adoption of the euro impact trade volatility. Along with discussing possible contributory variables including economic integration and currency stabilization programs, this section also evaluates the ramifications of these patterns.

3 Data and Methodology

Data used in this study was sourced from the World Bank's Data Bank, using four major indicators: Exports of goods and services (% of GDP), Imports of goods and services (% of GDP), Exports of goods and services (constant 2015 US$), and Imports of goods and services (constant 2015 US$). The selected indicators provide a relative measure of trade concerning national output, as well as an absolute measurement of trade flows adjusted for inflation. In constant US dollars in the year 2015, this study allows for the comparison of trade volumes over different periods by accounting for the effects of inflation. These would also be relative measures, especially of exports and imports as shares of GDP, to abstract away the differences in economic size and hence make the analysis independent of the magnitude of the economy. On the other hand, absolute measures—exports and imports in constant 2015 US dollars—give a clearer view of the trade flows in nominal terms, adjusted for inflation. Taken together, these indicators give a broad view of the trade dynamics of each country and their relationship with European Union membership and the adoption of the euro. Trade data was collected for each country from the year preceding their membership in the EU or adoption of the euro to the latest available after their accession into the EU or introduction of the euro. This longitudinal data set will allow the examination of changes in the volatility of trade over time and thus allow a before-and-after analysis. To maintain the integrity of the dataset, missing data points were interpolated where necessary to achieve consistent time series across all variables. Outliers were detected and withdrawn from the data set. In a paper devoted to measuring trade volatility, the outliers would pose a great deal of harm as they disproportionately affected the measure of volatility and thereby distorted it. Absent data and anomalous values were treated cautiously to reduce potential biases related to data outliers.

The present paper looks at two distinct country groups:

The first group analyses the effect of EU membership on trade volatility. It comprises the following countries: Austria, Croatia, Cyprus, Czechia, Bulgaria, Romania, Poland, Estonia, Hungary, Latvia, Lithuania, Malta, Slovenia, the Slovak Republic, Sweden, and Finland (see Figure 1).

The next group of countries looks at the impact of euro adoption on trade volatility and includes Austria, Cyprus, Estonia, Finland, Latvia, Lithuania, Malta, Slovenia, and the Slovak Republic (Figure 2). The EU membership group contains countries that joined the EU at different times, while the euro adoption group looks at those that have adopted the euro. The aim is to find out how the volatility of trade develops when countries take these major steps in integrating their economies.

The measure covers calculations of standard deviations for exports and imports as a share of GDP and as a constant of USD 2015, separately for the periods before and after EU accession and euro adoption. Finally, percentage changes in trade metrics were calculated using the formula x/y−1, where x represents the post-adoption value, and represents the pre-adoption value. This formula expresses the proportional increase or decrease in trade activity, offering a clear picture of how trade flows evolve following integration into the EU or adoption of the euro through the use of both descriptive and inferential statistical analyses, this research explores the patterns related to trade volatility while trying to establish a correlation with broader economic indicators, hence giving a wide perspective of the trade dynamics within the framework of EU membership and the eurozone.

Standard deviation is one of the widely used measures of volatility and indicates the spread of trade flows around the average and thus a level of trade stability or instability. In addition to standard deviations, a set of descriptive statistics was calculated including mean, median, standard deviation, minimum, and maximum trade volatility for both periods—before and after EU membership or euro adoption. This battery of measures allows a thorough understanding of both the central tendency and dispersion of the trade volatility data.

Furthermore, this paper examines the relationship between trade volatility and recent GDP per capita levels, measured in constant 2015 US dollars. Correlation analysis was performed to assess the degree to which higher levels of economic development, as measured by GDP per capita, are associated with greater trade stability.

Table 1: European United Calculations

Country	Exports (% GDP)	Exports (constant USD)	Imports (% GDP)	Imports (constant USD)	GDP per capita (most recent year)
Austria	1.698	2.196	1.292	1.727	56,856.12
Croatia	0.623	0.614	0.612	0.31	21,460
Czechia	0.668	2.021	0.055	1.352	30,427.42
Cyprus	0.495	0.836	0.916	1	34,701.44
Bulgaria	0.368	-0.736	-0.287	-0.747	15,797.60

Romania	0.978	2.874	-0.0735	2.569	18,419.42
Poland	1.387	3.457	0.238	2.698	22,112.86
Estonia	0.603	3	0.0345	2.432	29,823.75
Hungary	-0.48	1.399	0.587	1.015	22,147.21
Latvia	4.103	3.35	0.96482	2.191	23,184.31
Lithuania	1.285	3.945	0.97	2.924	27,102.78
Malta	-0.105	21.447	-0.266	13.984	37,882.27
Slovenia	0.198	2.15	0.444	1.125	32,163.51
Slovak Rep.	-0.072	3.133	-0.084	-0.084	24,470.24
Sweden	-0.141	2.287	0.0722	2.174	56,305.25
Finland	-0.199	2.118	1.004	2.666	53,755.91
Average	0.7130625	3.3806875	0.40493875	2.3335	31,663.12
Median	0.549	2.2415	0.341	1.9505	28,463.27
Std. Dev.	1.096590684	4.964819026	0.5107126859	3.288097991	13267.72009
Minimum	-0.48	-0.736	-0.287	-0.747	15,797.60
Maximum	4.103	21.447	1.292	13.984	56,856.12
Correlation	-0.1834440382	0.1412313652	0.3655353612	0.2266844015	

Table 2: Eurozone Calculations

Country	Exports (% GDP)	Exports (constant USD)	Imports (% GDP)	Imports (constant USD)	GDP per capita (most recent year)
Austria	0.704	1.515	0.703	1.265	56856.12
Cyprus	0.641	0.716	1.018	0.778	34701.44
Estonia	0.0798	-0.106	-0.108	0.039	29823.75
Latvia	-0.553	-0.358	-0.295	-0.229	23184.31
Lithuania	-0.5756	0.146	-0.362	-0.098	27102.78
Malta	-0.684	4.66	-0.665	3.749	37882.27
Slovenia	0.062	0.709	0.177	0.3	32163.51
Slovak Rep.	-0.459	0.032	-0.365	-0.079	24470.24
Finland	-0.497	-0.232	0.447	0.235	53755.91
Average	-0.1424222222	0.7868888889	0.06111111111	0.6622222222	35548.92556
Median	-0.459	0.146	-0.108	0.235	32163.51
Std. Dev.	0.5358197686	1.567999477	0.5627213885	1.251434055	12159.61855
Minimum	-0.684	-0.358	-0.665	-0.229	23184.31
Maximum	0.704	4.66	1.018	3.749	56856.12
Correlation	0.3823603231	0.2581026559	0.5749188181	0.3500710199	

4 Results

When looking at the datasets, the apparent outcome is that there is a significant correlation between GDP-per-capita and trade volatility as seen through the computerized correlation calculation which was positive for indicators except for exports (% of GDP) of the EU. The negative correlation between exports as a percentage of GDP and trade volatility within the EU can be explained by several EU-specific factors. First, EU member states have well-established, diversified trade relationships both within the single market and globally, which helps to lessen exposure to market-specific fluctuations. Internal to the EU is its internal market: free movement of goods, services, capital, and labor; this smooths out disruptions to trade flows within member countries and reduces the volatilities associated with those disruptions. EU countries that are higher exporters of GDP generally have relatively competitive high-value export sectors, such as advanced machinery, pharmaceuticals, and technology, that tend not to fluctuate as wildly as commodity-based exports. The underlying strength of EU institutional bodies, like the European Central Bank and the European Commission, provides a framework for coordinating fiscal and monetary policy that cushions negative external shocks, adding stability to trade. An increased export-to-GDP ratio may often signal a situation in which EU member states are increasingly part of global value chains and thus much less susceptible to insulated disruptions. The combination of diversified and stable export sectors, strong institutions, and economic integration provides a comprehensive explanation of why higher export-to-GDP ratios can result in low trade volatility in EU countries.

The positive correlation between trade volatility and GDP per capita, especially the notably high correlation of 0.5749 with imports (% of GDP) in the Eurozone, suggests a range of factors reflecting the intricacies of economic integration and structural dynamics within the Eurozone. These findings are especially important when considering the EU and Eurozone datasets separately, as each offers unique insights into the broader patterns of trade volatility within the respective economic unions. The correlation in the Eurozone can be a result of greater openness of the economies, which naturally comes with removing trade barriers like tariffs and quotas-a consequence of EU and hence Eurozone membership. Being a part of such an economic union integrates the economies of the member states more deeply into the global economy, where trade flows are less constrained by domestic regulations. This increased openness can dramatically widen the exposure to international economic cycles, such as changes in demand for goods and services, financial crises, and sector-specific shocks in trading partners. Any significant change in this global economy will thus have a highly pronounced impact on trade flows in the Eurozone countries and hence a higher volatility of imports in terms of GDP. The relatively high sensitivity to external factors thus reflects greater integration with international trade dynamics of the Eurozone.

Moreover, comparative advantage-based specialization within the Eurozone may also serve to heighten the volatility of trade, especially in those industries where countries decide to concentrate their economic activities. Specialization contributes to the fact that countries of the Eurozone, while being highly competitive in a particular industry, are more susceptible to

industry-specific shocks. For example, in countries whose export base is highly concentrated in one or a few commodities or products, a change in world demand for that commodity will result in larger changes to trade patterns. Dependence on a narrow range of industries or products increases vulnerability to world price changes and economic shocks, increasing volatility in trade flows.

Another mechanism through which trade volatility arises is capital flow integration, as seen within the Eurozone. Eurozone membership promotes free movement of capital across member states hence stimulating economic growth and expansion of trade. On the other hand, integration of capital markets increases the economies' vulnerability to external financial crises since capital can quickly shift due to changes in the global economic environment. The high volatility of imports, which shares a high value of correlation in the Eurozone area, can be related to the fact that financial crises or sudden changes in investment sentiments abruptly alter the demand for imports and increase trade volatility.

Trade volatility could also increase because of the rigid exchange rate regime enforced through the use of a single currency, the euro. Individual economies of the eurozone no longer can devalue their national currencies in the case of external shocks. While the euro removes the risk of exchange rate fluctuations within the Eurozone, it also removes the flexibility afforded by earlier exchange rate regimes. As a result, Eurozone economies become highly vulnerable when faced with asymmetric shock-for instance, a situation when one region witnesses a recession while other member states remain in growth there is limited currency devaluation as an instrument for adjusting trade imbalance situations, a potential cause for accentuated volatility, especially for imports.

Trade policy, as under the EU, that integrates uniform trade regulation and agreements also can be another contributor to heightened volatility. These converging policies could be disruptive to traditional trading relationships and trade flows, especially in the first years of membership, as countries get used to new trade agreements, external tariffs, and regulatory frameworks. These adjustments could lead to short-term disruptions that contribute to volatility in imports, especially where particular sectors are facing unexpected challenges or transitions in response to these policy changes. This in turn also amplifies trade volatility, while access to global markets and supply chains is facilitated by EU and Eurozone membership. Membership in the Eurozone allowed member countries to integrate deeper with global supply chains, which helped them export and import goods and services more efficiently. But the integration with global supply chains also makes a country vulnerable to disruptions in global supply chains, fluctuations in demand, and localized crises occurring outside the region. For example, natural disasters or geopolitical conflicts that disrupt important global supply chains can sharply alter the availability and price of imports, increasing trade volatility. Second, some of the observed volatility is due to measurement effects, particularly when imports are expressed as a percent of GDP. While trade volatility may therefore appear to increase even though the absolute fluctuation in imports does not change, the relative share of imports in GDP increases as trade volumes go up following membership.

The reason imports are so volatile about GDP may thus also be related to the increase in trade as a share of GDP after accession to the Eurozone. If trade is growing faster than GDP, the relative volatility of imports would appear magnified, even though the overall fluctuations in trade may not have changed significantly. Last, policy and institutional changes accompanying accession to EU and Eurozone membership may also create temporary trade disruptions and heightened volatility. Many economies, upon integrating into the EU's regulatory environment and adopting the euro, go through structural adjustments in their domestic economies. These adjustments may cause short-term instability that may be manifested as heightened volatility of imports while businesses and markets adjust to new economic frameworks.

Therefore, it follows that high trade volatility in imports as a percentage of GDP with GDP per capita in the Eurozone could reflect an increasing economic openness, industrial specialization, integration of capital flows, rigid exchange rate, and also the harmonization of trade policy coupled with deeper participation in global supply chains. These dynamics collectively increase exposure to exogenous shocks and fluctuations, leading to higher trade volatility in the Eurozone. The findings bring out the complex trade-offs that come with deeper economic integration and highlight the challenges that the countries of the Eurozone face in managing trade volatility in a highly interconnected global economy.

5 Conclusion

Using GDP per capita data and imports and exports as percentages of GDP and in constant 2015 US dollars, this study examined the impact of EU membership and the use of the euro currency on trade volatility to find trends and connections. The results show that trade volatility is increased by EU membership and adoption of the euro, especially for imports (as a percentage of GDP) within the eurozone. These findings highlight the double character of economic integration: it increases economic interdependence and facilitates cross-border transactions, but it also increases vulnerability to internal market fluctuations and external economic shocks.

These results have wider implications for the literature on trade dynamics and economic integration that go beyond the purview of this investigation. They support continuing discussions over the compromises involved in regional unification, especially in the way that trade stability is affected by structural and policy changes. This article does not, however, identify the precise methods or causes of the observed rise in trade volatility, which is one of its drawbacks. According to the analysis, market specialization, trade policy changes, and exchange rate stabilization might all play a role, but more investigation is required to confirm these theories.

In order to identify the precise causes of elevated trade volatility, future research could expand on this study by using more detailed data, such as firm-level studies or sector-specific trade trends. Furthermore, comparing the reported volatility to that of non-EU or non-eurozone nations may shed light on whether it is specific to these areas or a component of a larger global trend. Furthermore, comparing the reported volatility to that of non-EU or non-eurozone nations may shed light on whether it is specific to these areas or a component of a larger global trend. A

deeper comprehension of the long-term impacts of EU membership and euro adoption may also be possible by extending the temporal scope to encompass pre-accession and post-accession trade trends over a longer time. In addition to addressing this paper's shortcomings, these improvements would enhance knowledge of the intricate relationship between trade dynamics and economic integration.

Works Cited

Buti, Marco, and Paul van den Noord. "THE EURO: PAST SUCCESSES AND NEW CHALLENGES." National Institute Economic Review, no. 208, 2009, pp. 68–85. JSTOR, http://www.jstor.org/stable/23880760. Accessed 8 Nov. 2024.

Brunnermeier, Markus K., and Jean-Pierre Landau. The Euro at 20: Resilience and Fragility of the European Monetary Union. European Parliament, Policy Department for Economic, Scientific, and Quality of Life Policies, 2022, https://www.europarl.europa.eu/RegData/etudes/STUD/2022/703337/IPOL_STU(2022)703337_EN.pdf. Accessed 2 Feb. 2025.

Cavallo, Alberto, et al. "The Price Impact of Joining a Currency Union: Evidence from Latvia." IMF Economic Review, vol. 63, no. 2, 2015, pp. 281–97. JSTOR, http://www.jstor.org/stable/24738133. Accessed 10 Nov. 2024.

Claeys, Grégory, and Guntram B. Wolff. "Is the COVID-19 Crisis an Opportunity to Boost the Euro as a Global Currency?" Bruegel, 2020. JSTOR, http://www.jstor.org/stable/resrep28501. Accessed 11 Nov. 2024.

Darvas, Zsolt. "The Euro Crisis: Mission Accomplished?" World Policy Journal, vol. 30, no. 1, 2013, pp. 87–94. JSTOR, http://www.jstor.org/stable/43290399. Accessed 10 Nov. 2024.

Feldstein, Martin. "The Failure of the Euro: The Little Currency That Couldn't." Foreign Affairs, vol. 91, no. 1, 2012, pp. 105–16. JSTOR, http://www.jstor.org/stable/23217153. Accessed 8 Nov. 2024.

Ioannatos, Petros E. "Has the Euro Promoted Eurozone's Growth?" Journal of Economic Integration, vol. 33, no. 2, 2018, pp. 1388–411. JSTOR, http://www.jstor.org/stable/26431812. Accessed 5 Nov. 2024.

Ioannatos, Petros E. "Has the Euro Promoted Eurozone's Growth?" Journal of Economic Integration, vol. 33, no. 2, 2018, pp. 1388–411. JSTOR, http://www.jstor.org/stable/26431812. Accessed 11 Nov. 2024.

Kalaitzoglou, Iordanis, and Beatrice Durgheu. "Financial and Economic Growth in Europe: Is the Euro Beneficial for All Countries?" Journal of Economic Integration, vol. 31, no. 2, 2016, pp. 414–71. JSTOR, http://www.jstor.org/stable/43783272. Accessed 11 Nov. 2024.

Lane, Philip R. "The Resilience of the Euro." The Journal of Economic Perspectives, vol. 35, no. 2, 2021, pp. 3–22. JSTOR, https://www.jstor.org/stable/27008027. Accessed 11 Nov. 2024.

Norrlof, Carla. "Key Currency Competition: The Euro versus the Dollar." Cooperation and Conflict, vol. 44, no. 4, 2009, pp. 420–42. JSTOR, http://www.jstor.org/stable/45084584. Accessed 8 Nov. 2024.

Rich, Georg. "The Euro After Five Years." The Brown Journal of World Affairs, vol. 11, no. 1, 2004, pp. 241–51. JSTOR, http://www.jstor.org/stable/24590513. Accessed 5 Nov. 2024.

Smith, Roy C. "The Agony of the Euro." The Independent Review, vol. 18, no. 1, 2013, pp. 49–76. JSTOR, http://www.jstor.org/stable/24563194. Accessed 10 Nov. 2024.

Silvia, Stephen J. "Is the Euro Working? The Euro and European Labour Markets." Journal of Public Policy, vol. 24, no. 2, 2004, pp. 147–68. JSTOR, http://www.jstor.org/stable/4007858. Accessed 5 Nov. 2024.

The International Role of the Euro. European Central Bank, 2015, https://www.ecb.europa.eu/pub/pdf/ire/euro-international-role-201507.en.pdf. Accessed 10 Nov. 2024.

AI-Driven Prediction: Enhancing Air Quality Prediction Using Longitude and Latitude
By Harshitha Sathyanarayanan

Abstract

The air humans rely on for breathing is being polluted by certain activities (e.g. burning of fossil fuels, vehicle emissions, agricultural activities, etc.), which transform a vital resource into a silent killer. Air pollution remains one of the most urgent and pressing environmental challenges, impacting the health of billions and accelerating climate change. As urbanization and industrial activity expand, the need for innovative solutions to monitor and predict air quality has never been more critical. Although certain technologies exist to monitor and predict air quality, they often lack accuracy and reliability. Traditional air quality monitoring systems are effective in specific areas but present challenges when expanded to larger regions. Their substantial size and weight require dedicated infrastructure, leading to higher installation and maintenance costs. This bulkiness limits the number of units that can be deployed, resulting in sparse coverage and an inability to capture the spatial variability of air pollution across diverse areas. Consequently, these systems often fail to provide comprehensive real-time data necessary for timely public health interventions. These technologies struggle to account for complex factors such as diverse pollution sources, unpredictable weather patterns, etc. This project seeks to address these limitations by investigating how effectively AI can predict AQI (Air Quality Index) trends over time in various regions with diverse climates and pollution sources. Historical and real-time air quality data was utilized to evaluate the performance of various AI models, identify limitations, and optimize accuracy. The methodology includes collecting and preprocessing datasets, training models to account for geographic and environmental differences, and analyzing results to propose enhancements for real-world applications. The results indicate that the KNN Classifier achieved the highest accuracy for classifying AQI based on geographic coordinates for PM2.5, while Logistic Regression and XGB Classifier excelled in classifying AQI for CO (Carbon Monoxide) and Ozone. The KNN Classifier also had the best overall performance, with the highest average accuracy and lowest standard deviation, suggesting minimal variability across different pollutants. While the results were promising, there is still potential for refinement, by incorporating more environmental factors and fine-tuning model parameters. Future work could expand the dataset and include real-time data, enhancing the model's accuracy and global applicability.

Introduction

The natural environment faces numerous threats, including climate change, habitat destruction, deforestation, and population growth. Among these challenges, air pollution is a significant concern, widely regarded as one of the most pressing issues affecting both ecological health and human well-being. Exposure to air pollutants causes environmental harm such as acid rain, which damages forests, degrades soil, and harms aquatic ecosystems (National Institute of Environmental Health Sciences, 2025). Air pollutants such as NO_2, CO, Ozone, SO_2, PM2.5, and

more pose serious risks to human health, often going unnoticed when inhaled. Once they enter the bloodstream, they can cause various illnesses, such as lung diseases, cancer, asthma, breathing problems, fatigue, and even premature death (European Environment Agency, 2024). The World Health Organization (WHO) identifies air pollution as one of the leading environmental risks, as it contributes to millions of premature deaths each year (WHO, 2024). According to WHO, exposure to certain pollutants like particulate matter (PM2.5) and nitrogen dioxide (NO_2) cause a range of respiratory and cardiovascular diseases. They emphasize the urgent need for reliable air quality monitoring systems (WHO, 2024).

Traditional monitoring systems can be effective in small, specific areas. However, they are far more costly to scale and often fail to provide complete real-time data across broader and more diverse regions.Traditional air quality monitoring systems are often bulky, and heavy, which make them inconvenient for deployment. For instance, the Continuous Ambient Air Quality Monitoring Station (CAAQMS) ranges from 9.8 kg to 35 kg (Prana Air, 2024). Traditional air quality monitoring systems are also costly, with initial installation expenses reaching thousands of dollars and ongoing maintenance adding to the financial burden. According to an article by Maximize Market Research, the installation of a single air quality monitoring station ranges from $100,000 to $200,000 (Maximize Market Research, 2024). Additionally, an article by FinModelsLab states that, " …the sensor equipment and its ongoing maintenance represent a significant portion of the operating costs…The initial capital outlay for the sensor equipment can range from $5,000 to $15,000 per monitoring station, depending on the sophistication of the sensors and the specific requirements of the project," (Sheykin, 2024). Because of their large sizes and high costs, these devices are sparsely deployed; this often leads to limited spatial coverage, meaning that the data that these devices provide don't accurately represent the air quality in surrounding areas.

Many concerns about air quality stem from human activities. According to an article by National Geographic, "Most air pollution is created by people burning fossil fuels, which include coal, natural gas and oil," (Rutledge et al., 2024). The burning of fossil fuels releases large quantities of harmful gases, including carbon dioxide (CO_2), nitrogen dioxide (NO_2), methane (CH_4), nitrous oxide (N_2O), and carbon monoxide (CO), among others. According to the EPA, the most common air pollutants in the air include Ozone (O_3), Particulate Matter (PM2.5), Carbon Monoxide (CO), Sulfur Dioxide (SO_2), Nitrogen Dioxide (NO_2), and Lead (Pb), with the exact concentration of each pollutant differing based on factors like urban density and industrial activity (EPA, 2021). These gases trap heat in the Earth's atmosphere, intensifying the greenhouse effect and accelerating global warming. The consequences of this include rising global temperatures, melting ice caps, more frequent extreme weather events, and disruptions to ecosystems and agriculture. Air pollution also leads to the formation of ground-level ozone, which can harm plant growth, reduce agricultural yields, and disrupt ecosystems. These impacts highlight the importance of addressing this pressing issue. Addressing air pollution and reducing fossil fuel consumption are essential steps toward mitigating the impacts of global warming and protecting the planet's future.

Current research highlights a significant gap in the availability of effective physical devices or instruments capable of accurately monitoring air pollutants in real time. According to Dr. William McCann, the chief medical officer of Allergy Partners, "there are two main types of air quality monitor sensors: particulate monitors that detect tiny motes (pieces) of dust, pollen, mold, and smoke, and gas monitors that can sense gases like carbon dioxide. The best air quality monitors will have multiple sensors so they can detect *both* particulate matter and gases…" (Laukkonen, 2024). However most air quality monitors are complex, and most sensors may not be able to detect both particulate matter and gases simultaneously due to technical limitations or design constraints. Additionally, existing technologies are often unreliable due to their high costs, impractical deployment, and occasional inaccuracies. For instance, a study published in the journal *Atmosphere* highlights that traditional regulatory-grade networks are limited by their sparse spatial coverage and high deployment costs, making them less effective in providing real-time data across diverse urban environments (De Vito et al., 2024). Air quality models have to consider a wide range of factors that constantly change and interact with each other, making them harder to create and use effectively. One major obstacle in predicting air quality is the unpredictability of natural disasters, which can dramatically alter pollutant levels in the atmosphere. According to the U.S. Geological Survey, "The 1980 eruption of Mount St. Helens vented approximately 10 million tons of CO2 into the atmosphere in only 9 hours," (Volcano Hazards Program). Another example is how carbon dioxide (CO_2) levels can rise sharply after a wildfire, making predictions inaccurate. Air quality is always changing, affected by factors like weather conditions (temperature, humidity, wind), emissions from human activities (factories and cars), natural events (such as wildfires), and chemical reactions in the atmosphere that can create new pollutants. Real-time sensors are useful but have limitations after wildfires and other disasters. They don't cover all areas, especially remote or heavily affected regions, and some data require processing before being useful. Air quality can change rapidly as smoke spreads with wind and weather, and sensors may not capture all harmful pollutants formed through chemical reactions. Models help fill these gaps by estimating conditions where sensors are lacking and forecasting how pollution will move.

The accuracy of air quality forecasts depends heavily on the quality and availability of data. Limitations in data collection infrastructure and inconsistencies in data availability pose significant challenges to developing accurate predictive models. A study published in *Atmosphere* highlights the challenges faced by low-cost air quality monitoring systems (LCAQMSs) in urban environments, emphasizing the need for accurate and personalized monitoring solutions. The authors stress that without advanced data management technologies and effective communication tools, these challenges can result in a lack of stakeholder trust, awareness, and environmental inequalities (De Vito et al., 2024). Overcoming these obstacles will require the use of advanced technologies, such as machine learning, to effectively process and analyze complex datasets, ultimately enabling more reliable and actionable air quality predictions.

As AI and machine learning become more regularly integrated into human life, it should also be used to solve environmental challenges. AI and machine learning models are becoming more widely used for predicting future events. According to an article by the Alexander Von Humboldt Institute for Internet and Society, "...it is increasingly being used to make predictions about future events on various social levels. From weather forecasting to the financial markets and medical diagnoses, AI systems promise more accurate predictions and improved decision-making," (Mosene, 2024). These technologies offer powerful methods for predicting air quality. Specifically, regression and classification methods are used in prediction models to analyze data and forecast outcomes. Regression helps predict continuous variables, such as pollutant levels, while classification categorizes data into different air quality levels. These methods analyze patterns in historical and real-time data to provide more accurate and reliable predictions, offering a valuable tool for addressing air pollution. This research aims to implement regression and classification techniques to analyze and predict air quality with greater precision. The primary objective is to enhance the understanding and reliability of air quality forecasting for addressing public health and environmental concerns.

Literature Review

Rapid economic growth, high-intensity pollutant emissions, and unfavorable weather conditions have significantly contributed to declining air quality in urban regions. For some context, air pollution is prominently worse in urban areas due to factors like increased vehicle traffic, industrial activity, and population density. Machine learning has been found as a promising tool for predicting air pollution and many studies have used it successfully. Research focusing on Zhangdian District utilizes a random forest algorithm to predict air quality by incorporating data relating to industrial waste gas emissions, meteorological factors, and chemical interactions (Liu et al., 2022). The study explored how adjusting daily industrial emissions based on meteorological forecasts could maintain air quality at acceptable levels. This highlights the potential of predictive models to guide both policy decisions and sustainable industrial practices. Another similar study was conducted using a dataset from Jinan China. This research introduced a new air quality prediction approach that made use of meteorological data and pollutant concentrations. Ten meteorological factors were selected, and a random forest approach was applied to assess their impact on pollutant concentrations (Liu et al., 2024). Seasonal analysis showed that temperature, humidity, and air pressure significantly influenced air quality. These studies show that machine learning models that make use of meteorological factors and pollutant emissions are efficient in predicting air pollution and monitoring air quality. Another study used the GC-DCRNN model, a deep learning approach that predicted short-term PM2.5 concentrations by modeling spatial relationships through geographic features and captured temporal patterns with a sequence-to-sequence architecture, improving accuracy by 5%-10% over baseline models (Lin et al., 2018). The model incorporated factors such as park area size, the number of factories within set distances, and other geographic elements influencing air quality. These studies demonstrate the effectiveness of advanced machine learning models in

improving air quality predictions by capturing both spatial and temporal factors, contributing to more accurate pollutant forecasting.

While existing research has successfully used machine learning models, including random forests, decision trees, and deep learning techniques, to predict air quality based on pollutant emissions and meteorological factors, this study differs by focusing on classification of pollutant concentrations using the AQI index. Regression techniques have shown strong potential in prior studies, but this study investigates the use of multiple classification models to classify AQI based on geographic factors such as latitude and longitude. The dataset used in this project contains pollutant concentration data over a variety of geographical locations, making it more efficient. Additionally, this study looks into an extensive number of models (eg. Logistic Regression, Random Forest, etc) to find the best performing classification model. The models used in other studies use more complex models that aren't accessible to everyone, but this study uses more common and accessible models, making it easier to replicate. Furthermore, many previous studies have limited classification to fewer AQI categories (eg. "Moderate" and "Unhealthy"), whereas this research will classify all six AQI index categories, providing a more comprehensive analysis of pollutant levels across varying conditions.

Methods

The approach for this project involves customizing existing predictive models through comprehensive data exploration and analysis using Python. Data exploration helps understand collected data, enabling effective model adjustments for improved air quality prediction. Python functions are used to analyze and visualize the data. The dataset used for this project is obtained from Kaggle, an open-source platform. The dataset consists of 14 columns and 16,695 rows. The dataset provides information on the concentrations of different atmospheric pollutants in different cities around the world, along with the longitude and latitude of those cities (kaggle.com).

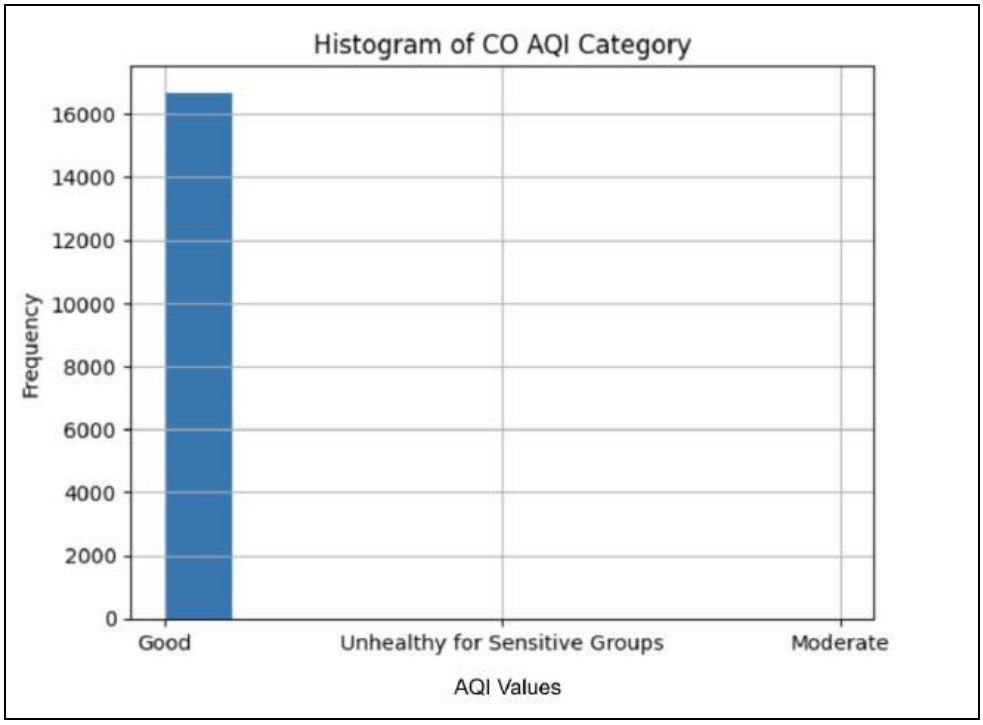


Figure 1: Histogram of Carbon Monoxide AQI Category Values

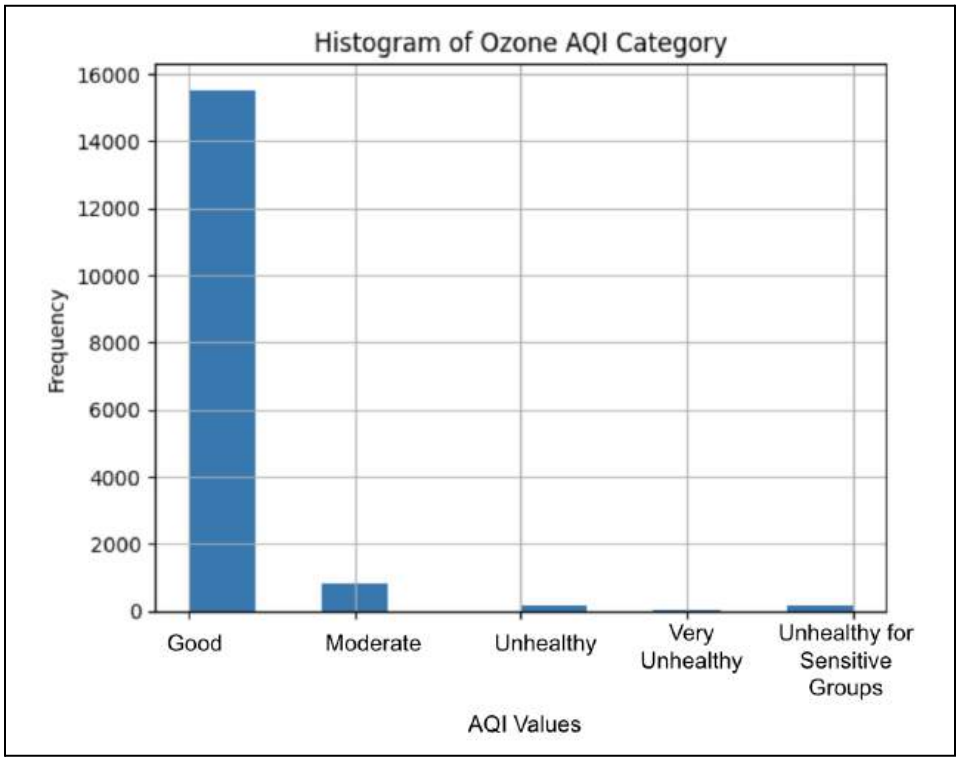


Figure 2: Histogram of Ozone AQI Category Values

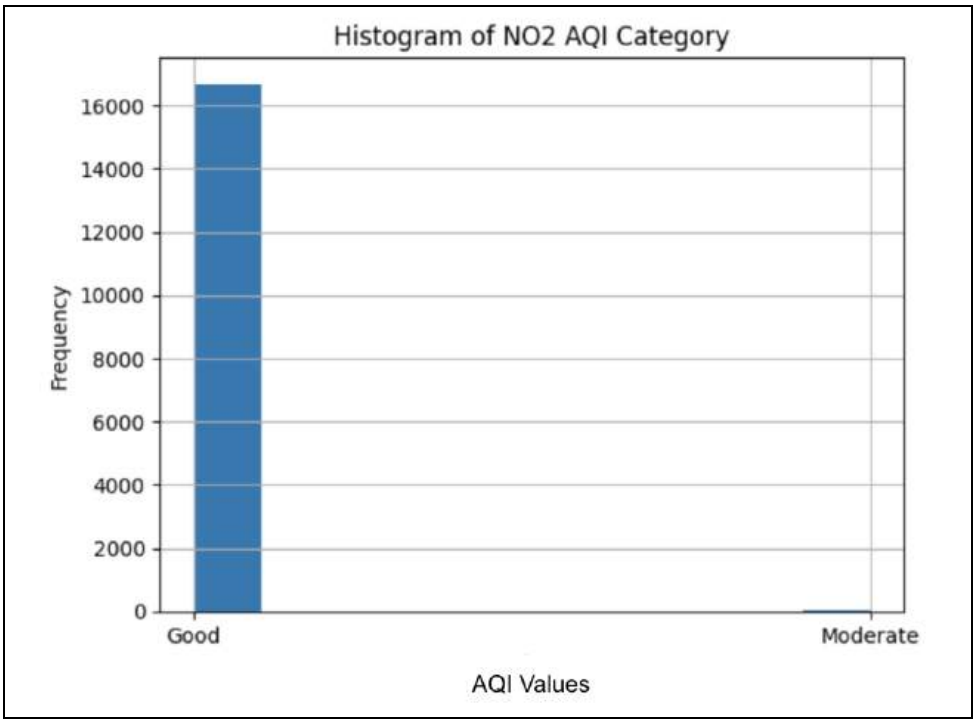


Figure 3: Histogram of Nitrogen Dioxide AQI Category Values

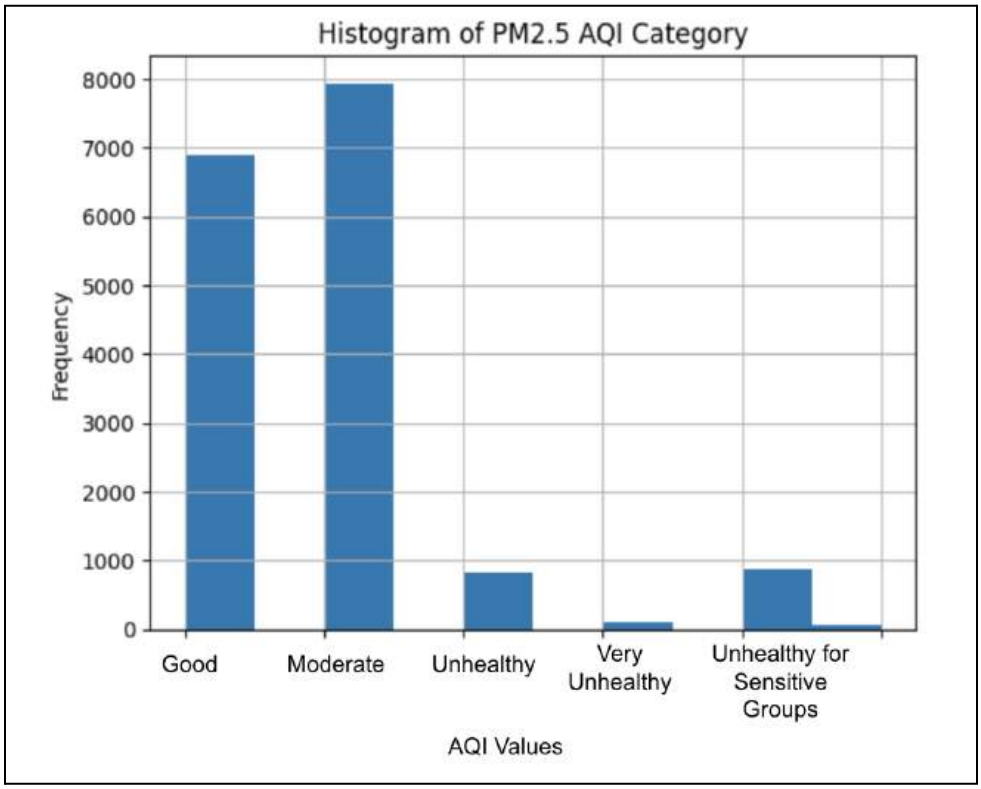


Figure 4: Histogram of Particulate Matter 2.5 AQI Category Values

The graphs shown above were generated in order to observe the different values stored under each categorical variable. In this case the different values under each categorical variable are the different AQI values. There are a total of six AQI values which have been determined by the EPA (Environmental Protection Agency): “Good”, “Moderate”, “Unhealthy”, “Very Unhealthy”, “Unhealthy for Sensitive Groups”, and “Hazardous”. In this dataset, the concentrations of CO in different cities fall under “Good”, “Unhealthy for Sensitive Groups”, or “Moderate”. Most of the different concentrations of CO for this dataset primarily fall under the AQI value “Good”. The Ozone concentrations in this dataset fall under “Good”, “Moderate”, “Unhealthy”, “Very Unhealthy”, or “Unhealthy for Sensitive Groups”. Most of the different concentrations of Ozone for this dataset also primarily fall under the AQI value “Good”. Most of the different concentrations of CO for this dataset primarily fall under the AQI value “Good”.The NO_2 concentrations either fall under “Good” or “Moderate”. Most of the different concentrations of NO_2 for this dataset primarily fall under the AQI value “Good”. The PM2.5 concentrations fall under “Good”, “Moderate”, “Unhealthy”, “Very Unhealthy” or “Unhealthy for Sensitive Groups”. Most of the different concentrations of PM2.5 for this dataset primarily fall under the AQI value “Moderate”.

The code for this project was implemented within a notebook on the Kaggle platform. The first step was importing libraries: NumPy for numerical computations and arrays, and Pandas for data analysis, cleaning, manipulation, and exploration. Matplotlib is known for creating visual and interactive models. For example, scatter plots, line plots, bar graphs, and histograms.The Sci-Kit Learn library provided most of the machine learning models used for this project such as Decision Trees, K-Nearest Neighbors, Naive Bayes, Random Forest, and Support Vector Classifiers, along with tools for standardization and model evaluation. Keras was used to build and train a neural network using a Sequential model with Dense layers. XGBoost, a machine learning algorithm known for its high performance in classification tasks, was also used.

A total of eight different classification models were tested in this process: Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Support Vector Machine Classifier, XGBoost Classifier, K-Nearest Neighbors, Naive Bayes Classifier, and the Artificial Neural Network Classifier. Logistic Regression is a linear model used for binary classification that predicts the probability of an outcome based on input features. Decision Tree Classifier splits data into branches based on feature values, making decisions through a tree-like structure. Random Forest Classifier combines multiple decision trees to improve accuracy and reduce overfitting of data. Support Vector Machine Classifier identifies the best way to separate different categories by drawing a boundary that maximizes the gap between the categories. XGBoost Classifier is an advanced boosting algorithm that improves prediction performance by refining decision trees. K-Nearest Neighbors determines the category of a data point by analyzing the classifications of its closest data points. Naive Bayes Classifier makes predictions using probabilities, assuming each piece of information contributes independently. Lastly, the

Artificial Neural Network Classifier mimics how the human brain processes information, using layers of artificial “neurons” to recognize complex patterns and make decisions.

Data cleaning and preprocessing are essential steps to ensure the dataset is accurate, consistent, and ready for analysis. The code reads the contents of the 'AQI and Lat Long of Countries.csv' file and stores it in the variable “df”, allowing for easy manipulation and analysis of the data. The code then removes the columns, “Country” and “City” from the dataframe. Since this project revolves around using numerical data to make predictions, there cannot be categorical data like country and city. Although categorical variables like “country” and “city” were dropped, there were still other categorical variables present in the dataset. However, those variables were necessary for the model to make predictions. In order for that data to be used in predictive modelling, it had to be converted into numerical values (integers). Each category within those columns was then represented by a new integer, depending on the number of classes belonging to that feature. A correlation study (Figure 5 shown below) was performed to analyze relationships between multiple air quality variables, including AQI values for pollutants like CO, Ozone, NO2, and PM2.5, as well as geographical coordinates (latitude and longitude) and AQI categories. The correlation matrix provides the correlation coefficients, indicating the strength and direction of associations between these variables.

	AQI Value	CO AQI Value	Ozone AQI Value	NO2 AQI Value	PM2.5 AQI Value	lat	lng	AQI Category_Good	AQI Category_Hazardous	AQI Category_Moderate	...
AQI Value	1.000000	0.458523	0.327956	0.308858	0.980360	-0.099554	0.150662	-0.558444	0.496177	0.048313	...
CO AQI Value	0.458523	1.000000	0.039752	0.399566	0.458846	-0.076412	0.002410	-0.213171	0.267346	0.000949	...
Ozone AQI Value	0.327956	0.039752	1.000000	-0.251169	0.233031	0.249740	0.318965	-0.120850	0.060555	-0.075776	...
NO2 AQI Value	0.308858	0.399566	-0.251169	1.000000	0.341455	0.040666	-0.290813	-0.313707	0.054454	0.160229	...
PM2.5 AQI Value	0.980360	0.458846	0.233031	0.341455	1.000000	-0.125997	0.113621	-0.588594	0.461840	0.089128	...
lat	-0.099554	-0.076412	0.249740	0.040666	-0.125997	1.000000	-0.027530	0.053613	-0.017943	0.039274	...
lng	0.150662	0.002410	0.318965	-0.290813	0.113621	-0.027530	1.000000	0.015529	0.046421	-0.133724	...
AQI Category_Good	-0.558444	-0.213171	-0.120850	-0.313707	-0.588594	0.053613	0.015529	1.000000	-0.056542	-0.792173	...
AQI Category_Hazardous	0.496177	0.267346	0.060555	0.054454	0.461840	-0.017943	0.046421	-0.056542	1.000000	-0.052224	...
AQI Category_Moderate	0.048313	0.000949	-0.075776	0.160229	0.089128	0.039274	-0.133724	-0.792173	-0.052224	1.000000	...
AQI Category_Unhealthy	0.561016	0.210598	0.324780	0.103662	0.551111	-0.098212	0.180469	-0.217277	-0.014324	-0.200682	...
AQI Category_Unhealthy for Sensitive Groups	0.311689	0.107652	0.024265	0.189829	0.313353	-0.081652	0.040897	-0.217014	-0.014307	-0.200438	...
AQI Category_Very Unhealthy	0.344155	0.213045	0.185672	0.098776	0.330458	-0.057332	0.071098	-0.082360	-0.005430	-0.076069	...
CO AQI Category_Good	-0.130860	-0.751786	-0.013884	-0.095904	-0.126988	0.002060	0.004526	0.014337	-0.189933	0.013242	...
CO AQI Category_Moderate	0.026224	0.214298	0.060061	-0.002395	0.017408	-0.008727	0.012497	-0.007168	-0.000473	-0.006620	...
CO AQI Category_Unhealthy for Sensitive Groups	0.135958	0.744329	-0.018647	0.112120	0.136578	0.002660	-0.012442	-0.012416	0.219582	-0.011467	...
Ozone AQI Category_Good	-0.353368	-0.089894	-0.717976	0.093383	-0.284079	-0.041335	-0.227842	0.253770	-0.072131	-0.055350	...
Ozone AQI Category_Moderate	0.167141	0.015465	0.336280	-0.094736	0.128734	0.057862	0.134595	-0.208585	0.045971	0.151863	...

Figure 5: Correlation Study

Before employing the models, the data set was split up into a training set and a testing set. To do this, certain libraries from Sci-Kit Learn were imported. The independent variable is set as the longitude and latitude coordinates, and the first dependent (or target) variable tested was the PM2.5 AQI Category. The models were trained using 70% of the dataset, and were tested using the remaining 30%. The code then adjusts the training and testing data to keep the feature values consistent and balanced, improving the model's performance. The Logistic Regression model was trained using the training data to learn patterns between inputs and targets, then tested on the testing set. Its accuracy was evaluated using the accuracy_score function, which compared predictions to actual results. This process was repeated for Decision Tree, Random Forest, ANN, SVM, XGB, KNN, and Naive Bayes models.

Results

Model:	**PM2.5**	**CO**	**Ozone**	**NO2**	**Standard Deviation**	**Average Model Prediction Accuracy Score**
Logistic Regression	0.52	1.00	0.93	1.00	**0.1996**	**0.8625**
Decision Tree Classifier	0.5626	0.9994	0.9323	0.999	**0.1814**	**0.8733**
Random Forest Classifier	0.6454	0.9998	0.9289	0.9992	**0.1460**	**0.8933**
SVM Classifier	0.5921	0.9998	0.9323	0.9992	**0.1690**	**0.881**
XGB Classifier	0.66	1.00	0.94	1.00	**0.14**	**0.9**
KNN Classifier	0.6606	0.9998	0.9351	0.9992	**0.1399**	**0.8986**
Naive Bayes Model	0.53	1.00	0.93	1.00	**0.1955**	**0.865**
Artificial Neural Network	0.0035	0.9994	0.9269	.9997	**0.422**	**0.7323**
Average Pollutant Prediction Accuracy Score	**0.5217**	**0.9997**	**0.9319**	**0.9995**		

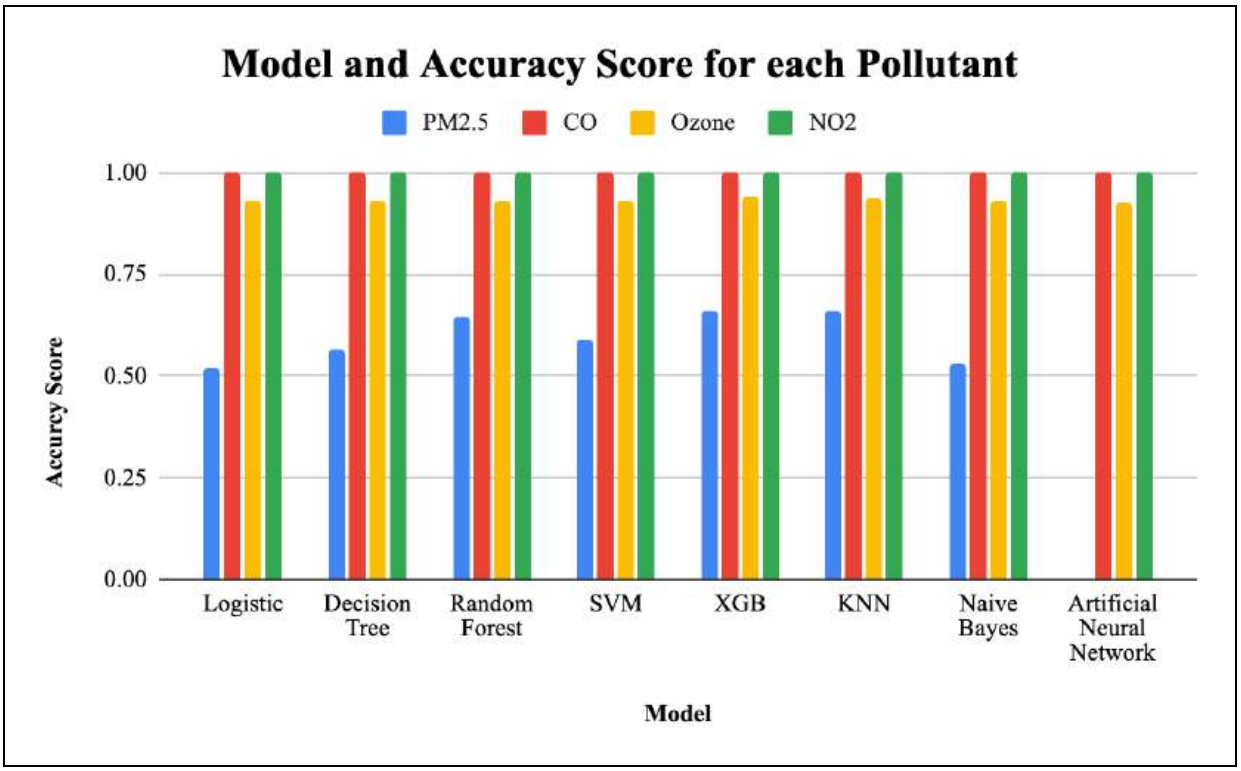


Figure 6: Bar Graph Showing Accuracy Score for Each Model and Pollutant

The KNN classifier achieved the highest accuracy for classifying AQI using latitude and longitude for PM2.5. For CO, Logistic Regression, XGB Classifier, and the Naive Bayes Model performed best. The XGB classifier showed the highest accuracy for Ozone, while Logistic Regression, the Naive Bayes Model, and XGB Classifier were most effective for NO2. Out of all the models tested, the XGB Classifier had the highest average accuracy score of 0.9, making it the most accurate model for classifying AQI based on longitude and latitude coordinates. The KNN Classifier had the lowest standard deviation of 0.1399, suggesting that the accuracy scores for the different pollutants are closely clustered around the average, with minimal variability among them. The logistic regression model had the highest standard deviation, meaning the accuracy scores are more spread apart (higher variability). Overall, the models were most accurate in classifying AQI for CO and NO2.

The correlation study performed earlier, showed varying relationships between pollutants, AQI values, and geographical features. The correlation between latitude and pollutants reveals distinct patterns. The correlation study reveals that latitude has the strongest positive correlation with Ozone AQI Value (0.25) compared to other pollutants, while longitude also shows a moderate correlation with Ozone (0.32). These geographic features likely contributed to the XGB Classifier performing best for Ozone AQI prediction, as the model effectively captured these spatial patterns. In contrast, PM2.5 showed a weaker correlation with latitude (-0.13) and longitude (0.11), yet the KNN Classifier achieved the highest accuracy for PM2.5, suggesting that localized patterns still played a role, aligning with KNN's strength in handling spatial

proximity. CO and NO2, which had lower correlations with latitude and longitude, saw Logistic Regression, Naive Bayes, and XGB performing best, indicating that non-spatial factors were more influential for predicting their AQI levels. These results suggest that pollutants with stronger geographic correlations benefit from models that can capture spatial trends more effectively.

Discussion

Classifying air quality using machine learning models is a challenging and vital task, as it provides a scalable and efficient way to monitor and predict air pollution, ultimately aiding in public health efforts, environmental protection, and policy decision-making. This study evaluated the performance of various machine learning models –Logistic Regression, Decision Tree Classifier, Random Forest Classifier, SVM Classifier, XGB Classifier, KNN Classifier, and Naive Bayes Model – for classifying air quality index (AQI) based on latitude and longitude coordinates across multiple pollutants (PM2.5, CO, Ozone, and NO_2). The KNN Classifier demonstrated the highest accuracy for classifying AQI for PM2.5, while Logistic Regression, XGB Classifier, and Naive Bayes models provided the best results for CO. For Ozone, XGB Classifier performed the best, and for NO2, both Logistic Regression and Naive Bayes were the most accurate. Among all tested models, the KNN Classifier achieved the highest average accuracy score of 0.8986, suggesting its prominent performance for AQI classification. It also had the lowest standard deviation of 0.1399, which indicates that its accuracy scores were consistently high and closely clustered around the average. In contrast, the Logistic Regression model had the highest standard deviation, suggesting greater variability in its results. Overall, this study found that KNN was the most reliable model for AQI classification, particularly for pollutants such as CO and NO2, where the models showed the most consistent performance.

Despite the promising results, there are several limitations in this study that could be addressed in future studies. Firstly, the process used in this study relied solely on latitude and longitude coordinates as features, which may notably capture the complexity of factors that influence air quality. Expanding the dataset to incorporate other relevant environmental factors, such as temperature, humidity, wind speed, or human activity, could improve model accuracy. Additionally, including more air quality features, such as concentrations of other pollutants (e.g. Particulate Matter 10, Sulfur Dioxide, or Lead), could lead to more accurate AQI classifications. The study also focused on specific pollutants and AQI data from a limited and unspecified time frame, which may have influenced the results. Broadening the dataset to include a variety of pollutants and data across different seasons, years, or time dimensions could provide a more comprehensive understanding of the models' performance over time.

Moving forward, there are several key steps that could significantly improve the model's performance and broaden its applicability. One potential direction is improving the feature engineering process to better capture the complexities of air quality data. For example, incorporating additional environmental variables, such as wind patterns, elevation, or atmospheric pressure could provide more context for predicting or classifying AQI. Additionally,

exploring alternative classification techniques, such as adjusting model parameters or trying different algorithms, could enhance prediction accuracy. For example, in the KNN Classifier, adjusting the number of neighbors (k-value) could help optimize the model for better accuracy. Similarly, other models like Logistic Regression or XGB Classifiers have certain parameters (such as learning rate and tree depth) which influence how well the model fits the data. By tuning these parameters, it is possible to find a more balanced model that performs better. Expanding the dataset could also improve the model's generalization. Including a broader range of geographical locations and diverse environmental conditions would allow for more accurate models that can handle variations in air quality patterns across different regions. Furthermore, incorporating real-time data (e.g. hourly AQI measurements or sensor-based readings), could provide a more up-to-date prediction. These models that are updated with new data over time would help maintain their accuracy, ensuring that they remain relevant as environmental conditions evolve.

Conclusion

This study showed that machine learning models can effectively classify AQI using pollutant concentrations and geographic data. The XGB Classifier had the highest overall accuracy, while the KNN Classifier was the most consistent. The hypothesis was only partially correct since KNN showed consistency but XGB provided the highest average accuracy. These results highlight the importance of AI in environmental monitoring by offering a scalable, data-driven method to assess air pollution. Accurate AQI classification can support public health efforts, guide pollution control strategies, and provide insights into pollutant behavior across regions. Future steps for this research include expanding the dataset to cover more regions and pollutants for broader applicability. Testing additional classification models and incorporating time-series data could further improve accuracy and consistency. Finally, integrating real-time data sources may enhance the model's practicality for continuous air quality monitoring and public health decision-making.

Works Cited

"Air Pollution and Your Health." *National Institute of Environmental Health Sciences*, U.S. Department of Health and Human Services, 27 Jan. 2025, www.niehs.nih.gov/health/topics/agents/air-pollution.

"Air Quality Monitoring System Market : IOT-Based Air Quality Monitoring System Has Provided a Better Approach towards Improving Environmental Conditions." *MAXIMIZE MARKET RESEARCH*, 7 Jan. 2024, www.maximizemarketresearch.com/market-report/air-quality-monitoring-system-market/162987/.

Alive, TK, et al. "How to Improve Classification Accuracy for Machine Learning." *Stack Overflow*, 1 Dec. 1961,stackoverflow.com/questions/41447104/how-to-improve-classification-accuracy-for-machine-learning.

"Ambient (Outdoor) Air Pollution." *World Health Organization*, World Health Organization, 24 Oct. 2024,www.who.int/news-room/fact-sheets/detail/ambient-%28 outdoor%29-air-quality-and-health.

Castelli, Mauro, et al. "A Machine Learning Approach to Predict Air Quality in California." *Wiley Online Library*, Wiley Online Library, 4 Aug. 2020, onlinelibrary.wiley.com/doi/full/10.1155/2020/6210201.

De Vito, Saverio, et al. "Future Low-Cost Urban Air Quality Monitoring Networks: Insights from the EU's AirHeritage Project." *MDPI*, Multidisciplinary Digital Publishing Institute, 10 Nov. 2024, www.mdpi.com/2073-4433/15/11/1351.

"How Air Pollution Affects Our Health." *European Environment Agency's Home Page*, 3 Dec. 2024, www.eea.europa.eu/en/topics/in-depth/air-pollution/eow-it-affects-our-health.

"Introduction to NumPy." *W3Schools*, www.w3schools.com/python/numpy/numpy_intro.asp#:~:text=What%20is%20NumPy%3F,%2C%20 fourier%20 transform%2C%20and%20 matrices Accessed 1 Feb. 2025.

Jain, Aarshay. "XGBOOST Parameters Tuning: A Complete Guide with Python Codes." *Analytics Vidhya*, 6 Jan. 2025, www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/.

Laukkonen, Jeremy. "Keep Your Eye on Pollutants with Our Favorite Air Quality Monitors." *The Spruce*, 30 Aug. 2024, www.thespruce.com/best-air-quality-monitors-4845803?utm_source=googlepaid&utm_medium=con&utm_content=Cj0KCQiAyoi8BhDvARIsAO_CDsDl0CEz6BsExE-k9RjM_tpK3thb81vCzIsb9RDKjeJ8uMUu3Fm_epwaAnFWEALw_wcB&utm_campaign=commerce-dd-AirQualityMonitors_TheSpruce_Combined_CommSEM_OrganicLP-4845803&utm_term=best+air+quality+monitor&utm_test=&displayPrice=yes&gad_source=1&gclid=Cj0KCQiAyoi8BhDvARIsAO_CDsDl0CEz6BsExE-k9RjM_tpK3thb81vCzIsb9RDKjeJ8uMUu3Fm_epwaAnFWEALw_wcB.

Lin, Yijun, et al. "Exploiting Spatiotemporal Patterns for Accurate Air Quality

Forecasting Using Deep Learning: Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems." *ACM Conferences*, 6 Nov. 2018, dl.acm.org/doi/abs/10.1145/3274895.3274907.

Liu, Qian, et al. "Air Quality Class Prediction Using Machine Learning Methods Based on Monitoring Data and Secondary Modeling." *MDPI*, Multidisciplinary Digital Publishing Institute, 30 Apr. 2024, www.mdpi.com/2073-4433/15/5/553.

Liu, Ying, et al. "Air Quality Prediction Models Based on Meteorological Factors and Real-Time Data of Industrial Waste Gas." *Nature News*, Nature Publishing Group, 3 June 2022, www.nature.com/articles/s41598-022-13579-2.

Natarajan, Suresh Kumar, et al. "Optimized Machine Learning Model for Air Quality Index Prediction in Major Cities in India." *Nature News*, Nature Publishing 'Group, 21 Mar. 2024, www.nature.com/articles/s41598-024-54807-1.

Noto, Joe Di. "How Much Does Air Quality Monitoring Cost?" *Kaiterra*, Kaiterra, 10 Jan. 2025, learn.kaiterra.com/en/resources/how-much-does-iaq-monitoring-cost.

"Our Nation's Air 2022." *EPA*, Environmental Protection Agency, gispub.epa.gov/air/trendsreport/2022/#home. Accessed 1 Feb. 2025.

"Prana Air CAAQMS Ambient Air Quality Monitoring Device: PM2.5, CO2, No2, SO2." *Prana Air*, 29 Oct. 2024, www.pranaair.com/us/air-quality-monitor/ambient-air-monitor/#:~:text=AQI%20Mobile%20App-,Dimensions:,%2C%20 OF%20 MOS%2C%20 Noise%2CCH.

Rutledge, Kim, et al. "Air Pollution." *National Geographic*, National Geographic Society, 15 Nov. 2024, education.nationalgeographic.org/resource/air-pollution/.

Sheykin, Henry. "How to Budget for Noise and Air Quality Monitoring?" *FinModelsLab*, 22 Nov. 2024, finmodelslab.com/blogs/operating-costs/noise-and-air-quality-monitoring-operating-costs

"Volcanoes Can Affect Climate." *USGS*, Volcano Hazards Program, www.usgs.gov/programs/VHP/volcanoes-can-affect-climate. Accessed 1 Feb. 2025

Yi, Wei Ying, et al. "A Survey of Wireless Sensor Network Based Air Pollution Monitoring Systems." *Sensors (Basel, Switzerland)*, U.S. National Library of Medicine, 12 Dec. 2015, pmc.ncbi.nlm.nih.gov/articles/PMC4721779/#:~:text=The%20drawbacks%20of%20the%20 conventional,low%20 spatial%20and%20 temporal%20 resolutions.

Zhang, Zhen, et al. "A Systematic Survey of Air Quality Prediction Based on Deep Learning." *Alexandria Engineering Journal*, Elsevier, 16 Mar. 2024, www.sciencedirect.com/science/article/pii/S1110016824002485#:~:text=Air%20q quality%20 prediction%20has%20 draws,prediction%20are%20 identified%20and%20 discussed.

Ziedler, Sarah. "One Step Forward, Two Steps Back: Why Artificial Intelligence Is Currently Mainly Predicting the Past." *HIIG*, 27 Nov. 2024, www.hiig.de/en/why-ai-is-currently-mainly-predicting-the-past/#:~:text=However%2C%20many%20 are%20 unaware%20that,predictions%20and%20 improved%20 decision%Making.

The Origin of the Nucleus and Endomembrane System in Eukaryogenesis: an Interdisciplinary Phylogenomic and Structural Perspective By Ayden Goh

Abstract

Among evolutionary problems, eukaryogenesis, the origin of the eukaryotic cell from its prokaryotic ancestors, is one of the most interesting. the defining organelle in a eukaryote is the nucleus, with many published theories proposing different evolutionary transition states to explain its origin. In particular, the syntrophy hypothesis proposes that the nucleus originated from an Asgard Archaeal endosymbiont that was taken up by a delta-proteobacteria host. By contrast, the Viral Eukaryogenesis (VE) theory proposes that the nucleus originated from a viral factory resulting from the infection of an Asgard Archaea by a virus within the *Tectillviricetes* family. This paper establishes a list of features that defined the Nucleus and associated Endomembranes of the Last Eukaryotic Common Ancestor (LECA). It then examines if the evolutionary transition states proposed by each hypothesis convincingly explains the development of the listed features in light of current Phylogenomic studies and structural analyses, with modifications to each theory proposed to address identified drawbacks. We find that the core idea behind the VE theory, the viral infection of an Asgard host, is more in line with current evidence than the core idea behind the syntrophy hypothesis. Our paper provides the evidence that support this perspective and provides new directions to test this hypothesis.

1.0 Introduction

Since the following structures were present in the Last Eukaryotic common ancestor (LECA)(Richards et al.), all hypotheses for the Eukaryogenesis must be able to explain the presence of the following features:

1. A nucleus containing a chimeric genome with genetic contributions from Asgard Archaea and alpha-proteobacteria (Richards et al.). In particular, the asgard contribution is dominant(*Dominant Contribution of Asgard Archaea to Eukaryogenesis | bioRxiv*).
2. The chimeric genome is organised in linear chromosomes.
3. A nuclear double membrane consisting of a bacterial type phospholipid bilayer as opposed to the archaeal monolayer of isoprene lipids(Zachar and Szathmáry).
4. Nuclear pores and customers involved in membrane bending. Nuclear pores and customers share a common ancestor, the proto coatomer. This is determined by observation that coatomer derived proteins involved in membrane bending comprise over half the mass of the nuclear pore complex (Field and Rout, “Coatomer in the Universe of Cellular Complexity”; Field and Rout, *Pore Timing*). The modern nucleocytoplasmic transport pathway is strongly associated with the Asgard genome (*Dominant Contribution of Asgard Archaea to Eukaryogenesis | bioRxiv*), so the proto coatomer should be of Asgard ancestry.
5. A differentiated endomembrane system with endoplasmic reticulum and Golgi body(Richards et al.).

This paper will outline the Viral Eukaryogenesis(VE) theory and the Syntrophy hypothesis, focusing on their proposals for Nuclear and endomembrane development. We then evaluate each theory's explanation for the presence of the five features listed above and propose modifications to correct identified shortcomings. Finally, we propose a synthesis of each theory with the proposed modifications.

While the original syntrophy hypothesis attempts to explain the formation of all of the aforementioned structures (except 2.), its proposed evolutionary transition states are either mechanically problematic (1. and 3.) or contradicted by primary research (4. and 5.). Moreover, it cannot be modified to explain 5. in a manner consistent with current research. The original VE theory completely omits any explanation for the origin of features 3. and 5., requiring modifications far more extensive than those proposed for the syntrophy hypothesis. Nevertheless, in its modified form, the VE theory explains the evolution of all 5 features. Thus, while we rate the original syntrophy hypothesis more highly than the original VE theory for attempting a more complete explanation of the origin of the endomembrane system, we present a modified VE theory which is more compelling than a revised syntrophy hypothesis.

2.0 Comparing the Viral Eukaryogenesis Theory and Syntrophy Hypothesis

2.1 Viral Eukaryogenesis Theory

The Viral Eukaryogenesis(VE) theory is divided into 4 phases. In the first phase, after attachment of the virus from the *Tectillviricetes* family to the Asgard host, a viral ortholog of the eukaryotic HAP2 fusogen facilitated the fusion of the Archaeal and viral internal membrane, releasing the viral linear DNA genome into the cytosol, where it lysogenised the host to form a virocell (Bell), establishing a protein bound viral factory similar to that in PhiKZ phages which eventually becomes the nucleus, immediately separating transcription ,occurring within the viral factory, from translation that occurs outside the viral factory. The fact that the viral factory was now filled with linear viral DNA forms the basis of the linear eukaryotic chromosome in which DNA is wound around histones. This is supported by research indicating that modern NCLDV viruses–with whom the *Tectillviricetes* family share a common ancestor–have their DNA packaged in eukaryotic-like nucleosomes (Liu et al.). The viral genome then encoded the enzymes that add the 5' m7G caps to mRNA to signal for its export into the cytoplasm for translation by host cell ribosomes(Bell). The virus also encoded a membrane bending protein meant for the formation of internal membranes of progeny virus, and this protein is proposed to be the common ancestor of modern eukaryotic nuclear pores and coatomer, the protocoatomer which both facilitated vesicle formation and was inserted into the protein barrier surrounding the viral factory as the ancestral nuclear pore (Bell). Here, the theory notes that these features lay the foundations for future endomembrane development (Bell).

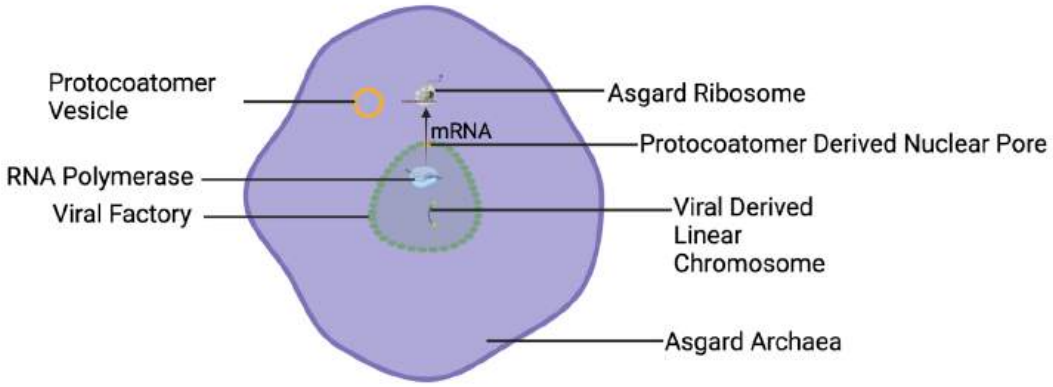


Fig 2.1.1: First Phase, post viral infection (Bell). The purple cell represents the Asgard host and the viral factory is in green. This infected cell is referred to as the virocell. The viral factory separates transcription within the viral factory from translation which occurs in the cytoplasm. The viral genome is contained within linear viral chromosomes. The viral protocoatomer facilitates the formation of both protocoatomer vesicles and protocoatomer derived nuclear pores(Bell).(Created using Biorender.com)

The second phase of eukaryogenesis involved the alpha-proteobacterial parasite entering the virocell, and its enslavement into an ATP exporting endosymbiont, while the third phase was characterised by the the transfer of alpha-proteobacterial and Asgard archaeal genes into the viral factory (Bell). Thus, the viral factory, at this point the nucleus, eventually gained the archaeal genes encoding the cellular translational system, as well as archaeal and alpha proteobacterial genes encoding basic metabolic pathways (Bell). This phase also involved the virocell, now the superorganism LECA, acquiring the ability to differentiate into motile flagellated or amoeboid forms, with a diverse range of eukaryotes evolving in the fourth phase (Bell).

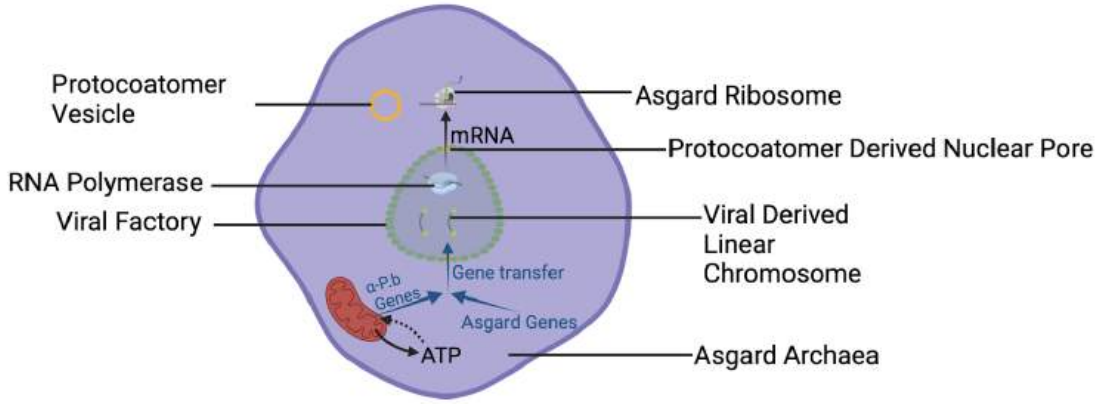


Fig 2.1.2: Second and Third Phases, before differentiation into flagellated and amoeboid forms(Bell). The red endosymbiont is the alpha-proteobacteria(mitochondria precursor). In the second phase, the alpha-proteobacteria parasite enters the cell by exploiting the vesicle system. It initially imports ATP from the host (dotted line) before being enslaved by the host to export ATP (solid line) infected cell is referred to as the virocell. In the third phase, there is a continuing extensive gene transfer from the alpha-proteobacteria and the Asgard genome into the viral factory(Bell).(Created using Biorender.com)

2.2 Syntrophy Hypothesis

The Syntrophy hypothesis is divided into 4 stages. It proposes that the nucleus and endomembrane system originated from the endosymbiosis of the Asgard archaea (and predates alpha-proteobacteria endosymbiosis), which initially existed in metabolic syntrophy with the delta-proteobacterial host (López-García and Moreira, "The Syntrophy Hypothesis for the Origin of Eukaryotes Revisited"). The theory proposes that complex fermentable organic substances were hydrolysed in the delta-proteobacterial periplasm before being transferred to the Asgard for further breakdown, and thus increasing the surface area of the periplasm in contact with the Asgard would facilitate the aforementioned movement of substances (López-García and Moreira, "The Syntrophy Hypothesis for the Origin of Eukaryotes Revisited"). Thus, at the first integration stage, the inner host membrane invaginated around Asgard (López-García and Moreira, "The Syntrophy Hypothesis for the Origin of Eukaryotes Revisited"). Gene transfer from the delta-proteobacteria host to the Asgard symbiont also begins in this phase (López-García and Moreira, "The Syntrophy Hypothesis for the Origin of Eukaryotes Revisited").

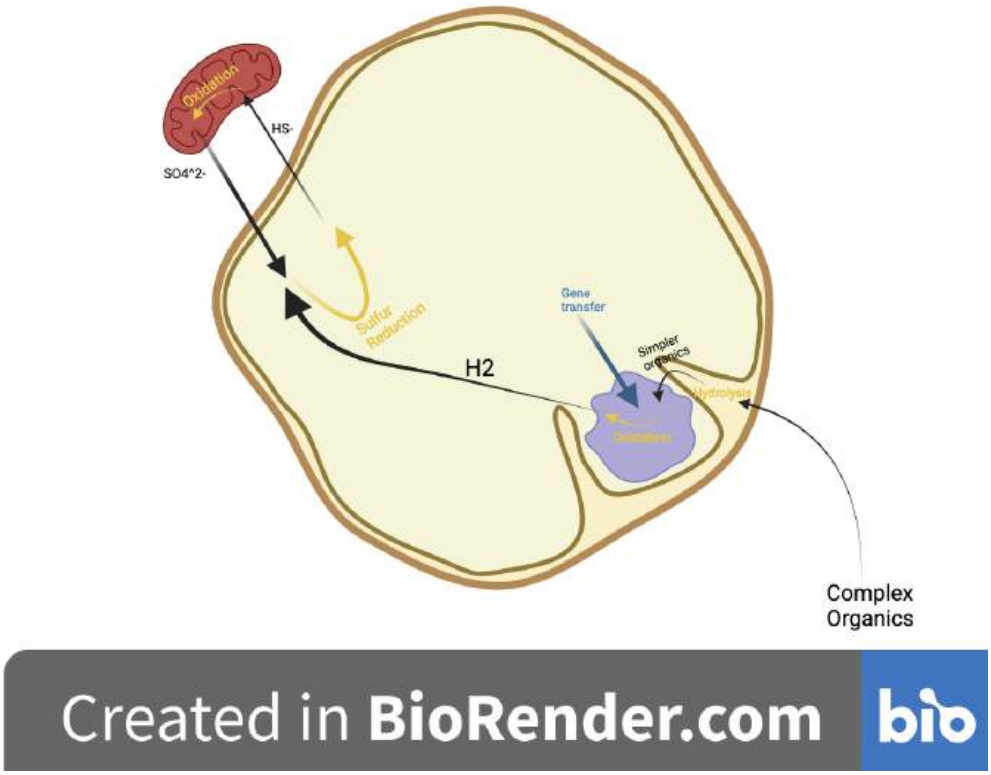


Fig 2.2.1: First Integration stage (López-García and Moreira, "The Syntrophy Hypothesis for the Origin of Eukaryotes Revisited"). The red cell represents the alpha-proteobacteria(mitochondria precursor), the purple cell the Asgard, and the yellow-brown cell the Delta-proteobacterial host. These three cells are collectively referred to as the consortium. Invangiation of inner delta-proteobacteria membrane enables periplasmic space to develop in Asgard's vicinity to facilitate transport of simple organics and amino acids (López-García and Moreira, "The Syntrophy Hypothesis for the Origin of Eukaryotes Revisited"). (Created using Biorender.com)

In the Second and Advanced integration steps, a second selection pressure now began to take precedence. The transfer of delta-proteobacteria genes from the host into the Asgard endosymbiont required the development of a transport system enabling the delta-proteobacterial gene products, especially the bacterial hydrolytic enzymes responsible for the aforementioned breakdown of complex carbohydrates in the periplasm, to be transported out of the Asgard through the Asgard's own membrane to their target areas (López-García and Moreira, "The Syntrophy Hypothesis for the Origin of Eukaryotes Revisited"). Moreover, since the transport system would be handling hydrolytic enzymes, it must have prevented their entry into the cytosol in order to prevent degradation of cytoplasmic components (López-García and Moreira, "The Syntrophy Hypothesis for the Origin of Eukaryotes Revisited"). And this selected for the development of a proto-endomembrane system. Which began with the bacterial proteins being inserted into the Asgard membrane in order to allow simpler bacterial gene products to pass through (López-García and Moreira, "The Syntrophy Hypothesis for the Origin of Eukaryotes Revisited"). This was then followed by the development of pore complexes on the bacterial membrane able to communicate with the Asgard membrane's transport proteins in order to allow more complex substances to be transported out of the Asgard proto-nucleus. A side effect of this was the congregation of ribosomes around the proto-nuclear pores in order to enable the translated gene products to be quickly exported (López-García and Moreira, "The Syntrophy Hypothesis for the Origin of Eukaryotes Revisited"). By contrast, transcription occurred primarily in Asgard's interior (López-García and Moreira, "The Syntrophy Hypothesis for the Origin of Eukaryotes Revisited").

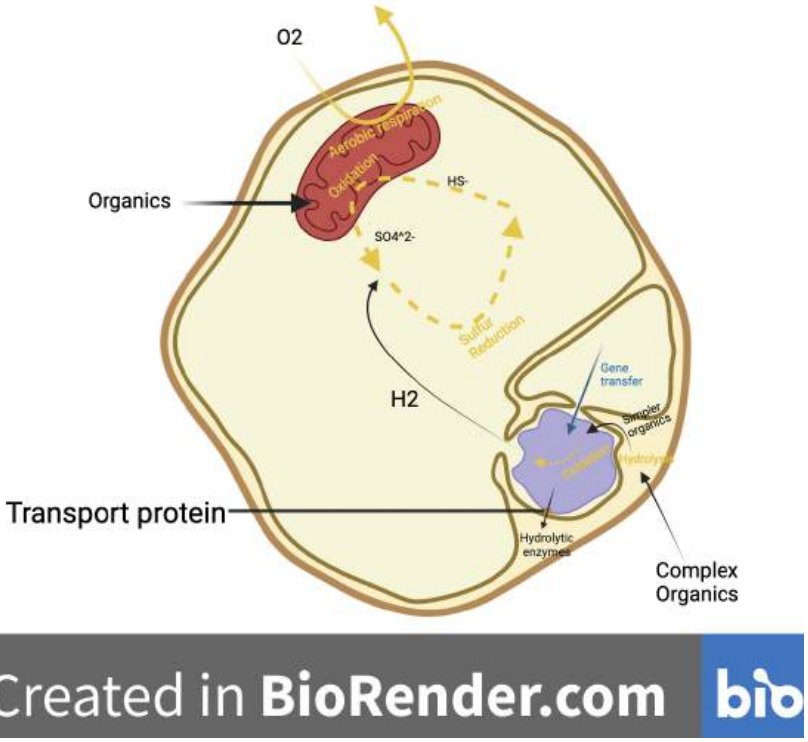


Fig 2.2.2: Second Integration step (López-García and Moreira, "The Syntrophy Hypothesis for the Origin of Eukaryotes Revisited"). Extensive development of the internal periplasmic membrane in contact with Asgard occurs to form the future nuclear membrane (López-García and Moreira, "The Syntrophy Hypothesis for the Origin of Eukaryotes Revisited"). Extensive gene transfer from the host to Asgard continues and Transport proteins are embedded in the Asgard membrane to allow delta proteobacterial gene products like hydrolytic enzymes to be exported out of the Asgard (López-García and Moreira, "The Syntrophy Hypothesis for the Origin of Eukaryotes Revisited"). Endomembrane system now takes on transport roles. (Created using Biorender.com)

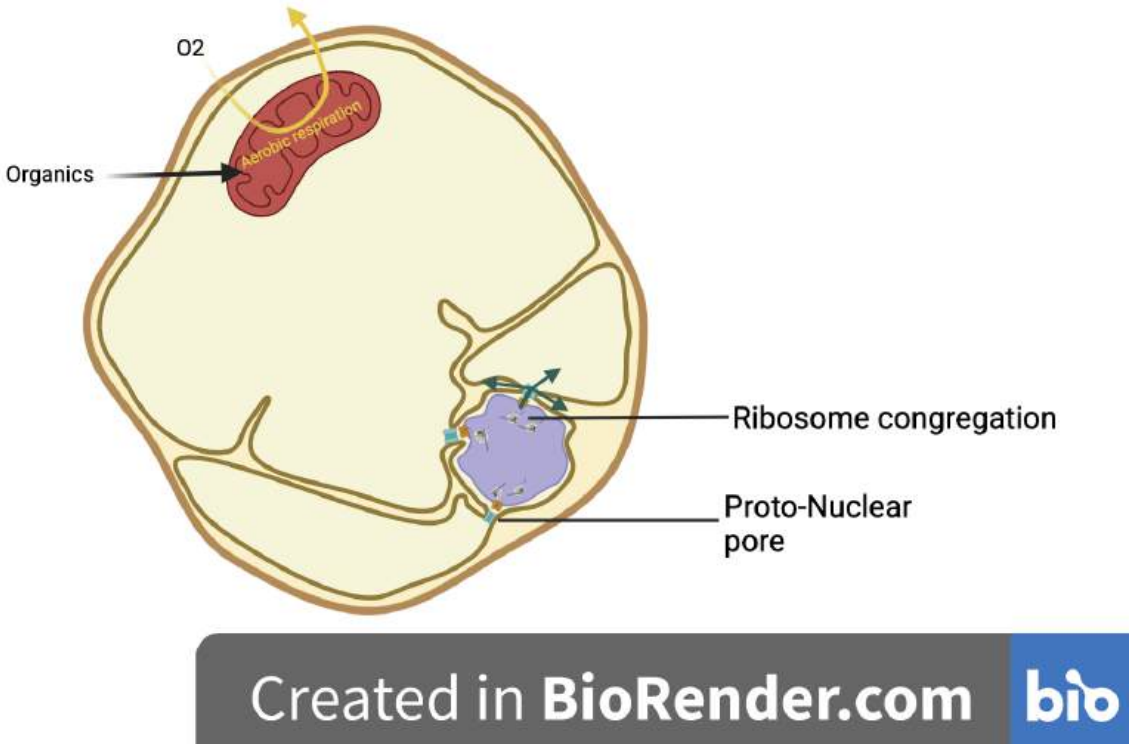


Fig 2.2.3 : Advanced integration stage (López-García and Moreira, "The Syntrophy Hypothesis for the Origin of Eukaryotes Revisited"). Blue-Green arrow represents export of Asgard gene products via endomembrane system. Proto nuclear pores (light blue-green) communicate with Asgard cell surface membrane transport proteins (orange) to facilitate export of complex gene products synthesised within Asgard. Ribosomes congregate near Asgard cell surface transport proteins communicating with proto-nuclear pores as translation locates preferentially to these areas (López-García and Moreira, "The Syntrophy Hypothesis for the Origin of Eukaryotes Revisited"). (Created using Biorender.com)

This spatial separation of transcription and translation proved invaluable in the LECA stage of eukaryogenesis . The loss of the Archaeal membrane allowed Mobile Group II introns from the alpha-proteobacterial endosymbiont to insert themselves through the host's genome, before fragmenting ,giving rise to spliceosomal introns and spliceosome encoding genes (Martin et al.; Martin and Koonin). This process is known as intron invasion. The far faster rate at which ribosomes translate mRNA compared to the rate of splicing required a clearer separation of translation from transcription and the attendant post-transcriptional modifications (splicing)

(López-García and Moreira, “The Syntrophy Hypothesis for the Origin of Eukaryotes Revisited”) than the aforementioned spatial separation. This third selection pressure then selected for consortia in which the former Asgard Ribosomes migrated out of the proto-nucleus (across the proto-nuclear pores) and eventually associated along the endomembrane system to form a proto rough endoplasmic reticulum on which translation took place (López-García and Moreira, “The Syntrophy Hypothesis for the Origin of Eukaryotes Revisited”). Meanwhile, the periplasm that once lined the delta-proteobacteria now became fully internalised (with some of its digestive functions now taken over by independent vesicles (López-García and Moreira, “The Syntrophy Hypothesis for the Origin of Eukaryotes Revisited”)), eventually becoming the cisternae lumen.

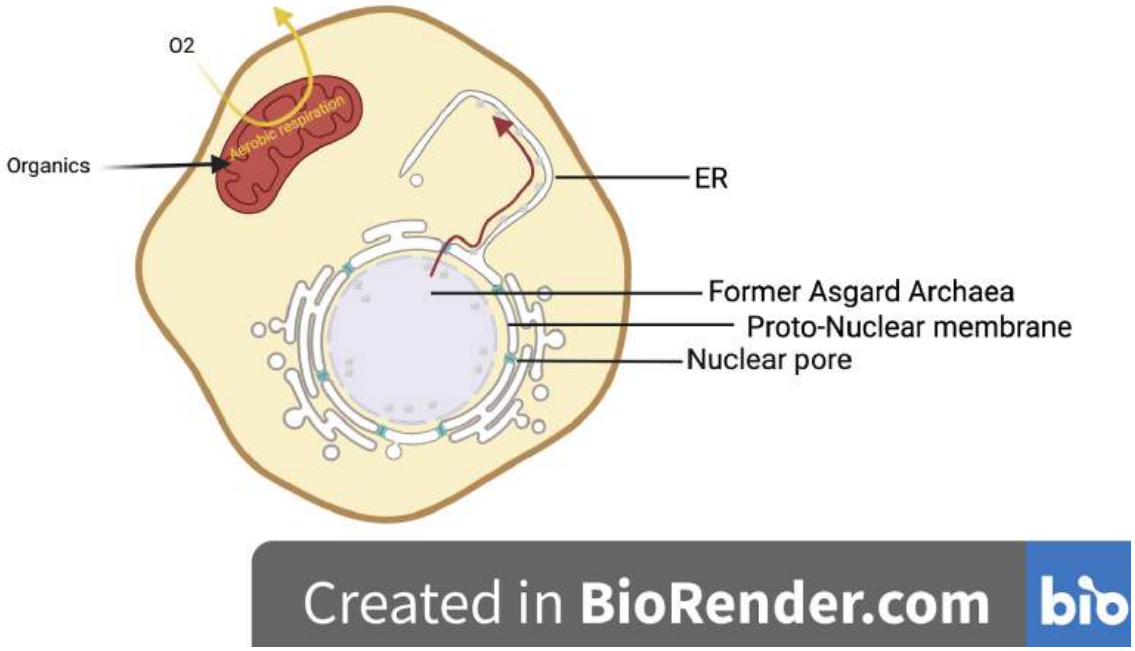


Fig 2.2.4: LECA stage (López-García and Moreira, “The Syntrophy Hypothesis for the Origin of Eukaryotes Revisited”). The Asgard Archaea’s membrane is now lost. The periplasm now becomes fully internalised as the endoplasmic reticulum and the red arrow represents the progressive movement of ribosomes out of the proro-nucleus along the endoplasmic reticulum. Intron invasion now results in selection for the proto-nuclear membrane to enforce decoupling of transcription and translation. Nuclear pore regulates traffic across the nuclear membrane and ribosomal subunits still assemble within the nucleus (López-García and Moreira, “The Syntrophy Hypothesis for the Origin of Eukaryotes Revisited”). (Created using Biorender.com)

3.0 Issues and Required Modifications

3.1 Gene Transfer resulting in the formation of the Chimeric Genome

The VE theory explains this well as is. It proposes that the chimeric genome forms when genes from both the host and the alpha-proteobacteria symbiont are transferred into the viral factory in the third phase(Bell). Since a proto-nuclear pore is introduced in the first phase (Bell), gene transfer into the viral factory, the proto-nucleus, is possible.

By contrast, this is impossible in the current Syntrophy hypothesis. It proposes that the chimeric genome is formed via gene transfer from both the delta-proteobacteria and alpha-proteobacteria into the Asgard(López-García and Moreira, "The Syntrophy Hypothesis for the Origin of Eukaryotes Revisited"). However, the existence of the Archaeal membrane presents difficulties for this (Jékely). Whereas gene transfer occurs most optimally when the symbiont lyses and releases its genome to integrate with the host genome (Martin and Russell), this method is clearly inapplicable in the opposite direction, as lysis of the host will kill the whole consortium (Jékely). And even if some other means of gene transfer was possible, such as the host's mRNA entering the asgard and being reverse transcribed, it would require the embedding of transport proteins within the asgard's membrane, in effect the development of a proto-nuclear pore complex, before gene transfer can begin (Jékely). In particular, the hypothesis's argument that gene transfer from the host to the symbiont is what eventually gives rise to the selection pressures that provide impetus for the evolution of the proto-nuclear pore complex (López-García and Moreira, "The Syntrophy Hypothesis for the Origin of Eukaryotes Revisited") is now unviable.

Nevertheless, we find this issue fixable, as a logical alternative selection pressure for the development of the proto-nuclear pore complex will be to facilitate this very gene transfer process.

3.2 Packaging of the Chimeric Genome in Linear Chromosomes

The VE theory explicitly addresses this issue by proposing that the Eukaryotic linear chromosome originated from the linear viral genomes in the viral factory(Bell).

While the syntrophy hypothesis does not explicitly address the development of linear chromosomes, the process of intron invasion it cites can explain the transition from circular to linear chromosomes. Specifically, recombination between 2 identical introns far apart from each other can cause the disintegration of the circular prokaryotic chromosome into linear chromosomes(Koonin).

3.3 Bacterial Type Nuclear Membrane

Both theories share a similar issue in that they propose that the proto-nucleus was originally bound by a non bacterial type membrane. Thus, this proto-nuclear membrane must be lost and replaced by a bacterial type nuclear envelope, which is an event the cell is unlikely to survive. The VE theory proposes that the nucleus was originally bound by a protein barrier analogous to that in PhiKZ phages to keep cellular nucleases out. The syntrophy hypothesis proposes that the inner delta-proteobacterial membrane invaginates around the Asgard archaea to form a bacterial type nuclear membrane when the Asgard membrane is eventually lost. However, these theories fail to resolve the timing and survival of the cell upon nuclear membrane replacement.

For the VE theory, the loss of the protein barrier would allow cellular nucleases to enter the nucleus. The Asgard archaea is predicted to have contained a variety of anti-viral defense

mechanisms from nucleases derived from CRISPR-Cas systems (Makarova et al.; Wu et al.) to programmable prokaryotic argonaute endonucleases, Restriction-Modification systems and viperins that convert nucleotides to nucleotide analogue inhibitors of viral DNA and RNA polymerases (Leão et al.). And since the genome is proposed to be reorganised along viral lines, with viral derived linear chromosomes, the only thing that prevented the genome from being degraded by the anti-viral mechanisms listed above was the protein barrier. This is supported by studies demonstrating that PhiKZ phage genomes are vulnerable to degredation by CRISPR Nucleases and Restriction-Modification enzyme systems that are able to get past the protein barrier(Mendoza et al.). Not only is there no selective advantage for the endomembrane to have developed around the viral nucleus to completely seal the viral genome from the cytoplasm so long as the protein barrier remained functional, but the moment even a section of protein barrier was rendered non-functional, cellular nucleases would have entered and degraded the genome within, likely killing the cell before the endomembrane could have evolved to seal the genome completely. To address this, we propose that gene transfer from to the viral genome, either from the host or from other organisms via horizontal gene transfer, enabled the viral genome to acquire genes that confer immunity from antiviral defense mechanisms within the same plasmid as other advantageous genes.

For the Syntrophy hypothesis,we propose to address this issue of cellular nucleases flooding into the nucleus upon archaeal membrane loss in a similar manner. The fact that the hypothesis proposes a delta-proteobacteria host that transfers genes to the Asgard archaea symbiont, combined with the fact that and bacterial plasmids encode genes that protect against cellular defense mechanisms like Restriction-Modification Systems (Dimitriu et al.) in order to facilitate conjugation, makes it entirely logical to propose that these plasmids were transferred into the Asgard genome during gene transfer, thus protecting the genome from degradation upon membrane loss. We also propose that other plasmids that confer protection against other defense mechanisms are transferred into the Asgard from other bacteria not part of the consortium.

However, the loss of the Asgard membrane presents a second issue unique to the syntrophy hypothesis, as it would result in the immediate loss of the archaeal membrane proteins that work in concert with the delta-proteobacterial proto-nuclear pores in the second and advanced integration stages(López-García and Moreira, “The Syntrophy Hypothesis for the Origin of Eukaryotes Revisited”). This would likely render the proto-nuclear pores non-functional. Thus, we propose that the proto-nuclear pores must be capable of functioning to the extent necessary for survival either with or without the associated Asgard membrane protein. Some degradation in the efficiency of the proto-nuclear pore is nonetheless expected, and we propose that this is selected for the replacement of the delta-proteobacterial proto-nuclear pore by an Asgard derived system, which is elaborated on in the next section.

3.4 Nuclear pores and membrane bending ability

Both theories share a similar shortcoming in that they propose that the nuclear pore and coatomer proteins conferring membrane bending ability originated from a non-Asgard source.

The Syntrophy hypothesis argues for a delta-proteobacterial origin of the nuclear pore and membrane bending ability. The VE theory argues for the viral origin of both. This is at odds with phylogenomic data that the nucleocytoplasmic transport pathway is strongly associated with the Asgard genome (*Dominant Contribution of Asgard Archaea to Eukaryogenesis | bioRxiv*). Combined with the fact that the Nuclear pore and membrane bending coatomers share a common ancestor known as the proto coatomer (Field and Rout, "Coatomer in the Universe of Cellular Complexity"), we conclude that both protein structures are of Asgard ancestry.

Uniquely for the syntrophy hypothesis, the nuclear pore consists of 2 different delta-proteobacterial proteins, one being inserted into the Asgard membrane and the other being inserted into the bacterial membrane. The former is considered in the next paragraph. For the latter, we propose it and the initial membrane bending proteins are initially of bacterial origin. This ensures that bacterial inner membrane invagination can take place as proposed in the hypothesis' first integration step before Asgard proteins can be exported into the bacteria. However, within the LECA stage after Asgard membrane loss, a second Asgard-derived proto coatomer is evolved and supplants its delta-proteobacterial equivalents, which are secondarily lost. Asgard protocoatomer derived nuclear pore proteins subsequently supplant the bacterial proto-nuclear pore, which is also secondarily lost. For a selection pressure to provide impetus for the above, we propose that Asgard derived proteins are more efficient than their delta-proteobacterial equivalents.

Although the protein inserted into the asgard membrane is eventually lost and is not the ancestor of the proto-nuclear pore, we conclude that it should still be of asgard origin. This is since transported proteins are always inserted from the cytoplasm into the cell surface membrane,so to ensure that the transport protein inserted into the Asgard membrane maintains the correct membrane topology, it must come from the Asgard Cytoplasm (Jékely).

For the VE theory, there is an additional issue of timing. It proposes that the proto-nuclear pore was formed before the differentiation of the primitive protocoatomer system. Structural analysis has revealed the nuclear pore is an amalgamation of both type I (typified by clathrin-coated vesicles and COPI transport vesicles) and type II (typified by COPII transport vesicles) coatomer proteins (Field and Rout, *Pore Timing*). Thus, the protocoatomer must have first differentiated into the aforementioned 2 major classes, type I and II. Since it is likely that this evolutionary change was selected for as it aided in the development of a differentiated endomembrane system, this indicates that the endomembrane development was at an advanced stage before the nuclear pore began to evolve(Field and Rout, *Pore Timing*). To address the theory's issues, we proposes that the viral proto-nuclear pore be embedded as proposed in the VE theory but is then secondarily lost sometime after the asgard derived protocoatomer is developed. The Asgard derived protocoatomer evolves in conjunction with the proposed development of the endomembrane system in section 3.5.

It is notable that the modifications proposed for both the syntrophy hypothesis and VE theory involve secondary loss of the non-Asgard proto-nuclear pore and replacement by Asgard protocoatomer. Further phylogenomic data to shed light on when nuclear pores developed, and

what role secondary loss played in it, would help determine if the author's modifications are viable. Alternatively, they would likely also provide clues on what other modifications can be made to address the issues identified.

3.5 Complex Endomembrane System

Both theories' explanations for the emergence of the eukaryotic complex endomembrane system fall short in different ways. For the VE theory, it simply fails to elucidate the evolution of differentiated structures within the endomembrane system beyond the protein-bound proto-nucleus and a primitive protocoatomer system even though LECA, which emerges in the third phase of the VE theory, is confirmed to have a complex endomembrane system (Richards et al.). For the Syntrophy Hypothesis, its proposed mechanism for endomembrane development is incompatible with recent phylogenomic research. In particular, it proposes the invagination of the inner delta-proteobacterial membrane and the transfer of relevant delta-proteobacterial to the Asgard's genome(López-García and Moreira, "The Syntrophy Hypothesis for the Origin of Eukaryotes Revisited"). This would imply a significant delta-proteobacterial contribution to the eukaryotic genome for genes associated with the endomembrane system, which is contradicted by the dominant Asgard contribution to the pathway for protein processing within the endoplasmic reticulum(*Dominant Contribution of Asgard Archaea to Eukaryogenesis | bioRxiv*).

For the VE theory, we propose that endomembrane development takes place in the third phase. Since this is when the superorganism gains the ability to differentiate into an amoeboid form capable of bacterial predation (Bell), the selection pressure for endomembrane development would then be to facilitate phagocytosis and breakdown of phagocytic products.Endomembrane development would follow three steps. In the first step, asgard derived vesicle formation from the cell surface membrane is evolved to facilitate phagocytosis. In the second step, translation of mRNA encoding hydrolytic enzymes in the cytosol would damage other cellular structures, the development of ribosomes that associate themselves with vesicles and thus release their hydrolytic gene product into the vesicle (proto lysosome) would then be selected for. This necessitates that Asgard genes encoding hydrolytic enzymes have been transferred into the viral factory. In the third step, the fact that it would be advantageous for these vesicles to be positioned close to the source of mRNA, the viral factory, leads to the amalgamation of such vesicles to form a continuous membrane bound body around the nucleus, which eventually differentiates to form the nuclear membrane endoplasmic reticulum and Golgi body. Additionally, the cell surface membrane must have changed from a monolayer of isoprene lipids characteristic of Archaea to a bacterial type phospholipid bilayer(Zachar and Szathmáry) before the above process. Since phylogenomic data does not offer substantial clarity regarding membrane replacement(*Dominant Contribution of Asgard Archaea to Eukaryogenesis | bioRxiv*) ,the author refrains from proposing a specific mechanism by which it occurred.

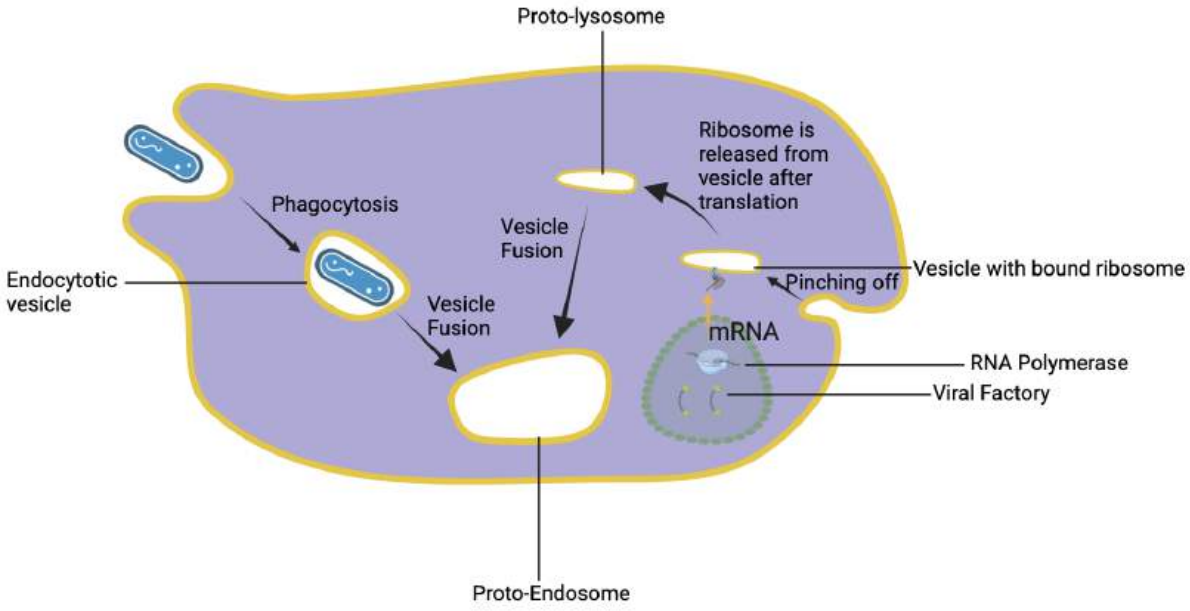


Fig 3.5.1: First and second steps of endomembrane development proposed for VE theory. Yellow membranes represent bacterial type phospholipid bilayers. The blue cell is the asgard host. Orange arrow represents movement of mRNA across the viral factory protein barrier. In the first step, Asgard vesicle coats evolve to facilitate membrane bending ability to facilitate phagocytosis. This also enables the pinching off of other vesicles from the cell surface membrane. In the second step, genes encoding hydrolytic enzymes are transferred to the viral factory. To prevent the release of hydrolytic enzymes into the cytoplasm, ribosomes associate with vesicles to translate mRNA encoding hydrolytic enzymes directly into the vesicle. (Created using Biorender.com)

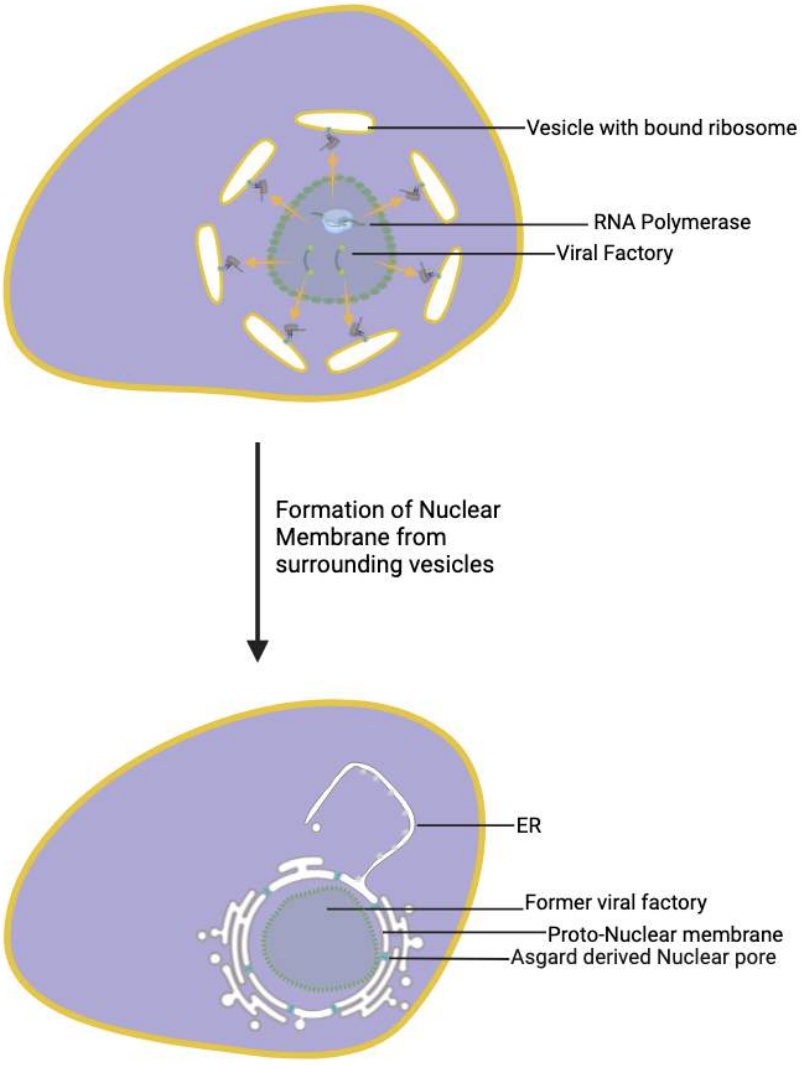


Fig 3.5.2: Third step of endomembrane development proposed for VE theory. Vesicles with associated ribosomes congregate around the source of mRNA, the viral factory. When the viral factory protein coat is lost, the congregation of vesicles surrounding the viral factory fuse to form the nuclear envelope and endoplasmic reticulum. Asgard derived nuclear pores embedded in the nuclear envelope, replacing the viral coatomer. genes encoding hydrolytic enzymes are transferred to the viral factory. (Created using Biorender.com)

By contrast, for the syntrophy hypothesis, we do not see how this issue can be remedied. This phylogenomic discrepancy is an inherent result of the delta-proteobacterial host that is central to the hypothesis. Furthermore, secondary loss of an entire delta-proteobacteria derived endomembrane system and replacement by another evolved from Asgard archaea would be unlikely.

The syntrophy hypothesis also neglects to explicitly elucidate the origin of the Golgi body. Nevertheless, we consider this issue to be minor as nothing in the theory precludes the

emergence of the Golgi body while the endoplasmic reticulum develops, since the eukaryotic membrane trafficking system likely evolved via the duplication and differentiation of the genes encoding other endomembrane organelles(Dacks and Field).

4.0 Proposed Synthesis of the Theories

In response to the drawbacks of both theories, we have proposed several revisions in section 3.0. In this section, we summarise the major changes along the timeline proposed by each hypothesis.

4.1 Modified Viral Eukaryogenesis Theory

Within the VE theory, the initial viral infection of the Asgard host is entirely plausible. This must have been followed by a change in the cell surface membrane from archaeal to bacterial in nature. This occurred between alpha-proteobacterial incorporation (second phase) and the differentiation into flagellated and amoeboid forms (late third phase). Simultaneously, gene transfer into the viral genome must have included the genes that protected the host from its own defense mechanisms against foreign mobile genetic elements and the genes encoding hydrolytic enzymes. Next, the endomembrane system evolved to facilitate phagocytosis, while the nuclear membrane originates from vesicles with bound ribosomes positioned as close as possible to the source of hydrolytic enzyme mRNA. This occurs in conjunction with the evolution of an Asgard derived protocoatomer which is more efficient than and thus replaces the viral protocoatomer, which is subsequently lost.

Our modified VE theory is compelling in several ways. First and foremost, it accounts for the evolution of the complex eukaryotic endomembrane system. Second, it defines a time by which membrane replacement must occur. Third, it accounts for the asgard origin of the modern eukaryotic protocoatomer system.

4.2 Modified Syntrophy Hypothesis

Proto-nuclear pore development is brought forward to the first integration step. In this, Asgard proteins were inserted into the Asgard membrane and delta-proteobacterial proteins were inserted into the bacterial membrane bounding the Asgard. Initially, these proteins work together as the initial proto-nuclear pore to facilitate gene transfer from the delta-proteobacteria host to the Asgard symbiont. Transferred genes include those encoding hydrolytic enzymes and those that confer protection from Asgard defense mechanisms against foreign mobile genetic elements. The latter can also be gained from bacteria outside the consortium by horizontal gene transfer. They then take on transport functions in the second integration step. After Asgard membrane loss in the LECA stage, the functionality of the remaining delta-proteobacterial half of the initial proto-nuclear pore was reduced to the level necessary for the organism to survive. An Asgard derived protocoatomer system evolves into both membrane bending vescicle coats and the nuclear pore, which replace their less efficient delta-proteobacterial counterparts. Endomembrane

development also includes the development of the Golgi body. Intron invasion results in the formation of the linear eukaryotic chromosome.

Our modified syntrophy hypothesis addresses criticisms of the mechanical impossibility of certain proposed evolutionary transitions and attempts to align the proposed evolutionary transition states with our current understanding of the molecular phylogeny of eukaryotes (*Dominant Contribution of Asgard Archaea to Eukaryogenesis | bioRxiv*). Nevertheless, particularly with regard to the origin of the eukaryotic endomembrane system, we were unsuccessful in explaining the proposed delta-proteobacterial origin of the endomembrane system, which must have entailed an outsized delta-proteobacterial contribution not currently observed in eukaryotic genomes(*Dominant Contribution of Asgard Archaea to Eukaryogenesis | bioRxiv*).

5.0 Conclusions

The original Syntrophy hypothesis attempts to explain the origin of most of the endomembrane system's structures. By contrast, the original VE theory completely fails to explain the origin of two structures–the nuclear envelope and the endomembrane system. Thus, the original syntrophy hypothesis is more commendable than the original VE theory

The original syntrophy hypothesis attempts but fails to address the origin of 4 of the 5 features, while the original VE theory only attempts to explain the origin of 3 of them. For the syntrophy hypothesis, the failure to address the 5th missing feature is only a technicality; the linear chromosome would result from intron invasion. For the VE theory, 2 features are not addressed at all; its proposed LECA still contained the viral proteinous nucleus and lacked differentiated endomembrane structures like the endoplasmic reticulum. We were thus required to synthesise from scratch an entire mechanism by which the endomembrane system evolved into the differentiated nucleus, endoplasmic reticulum and Golgi body.

Features of the Eukaryotic Endomembrane System	Syntrophy Hypothesis	VE Theory
Chimeric Genome Formed Via Gene Transfer	-/+	+/+
Linear Chromosomes	?/+	+/+
Bacterial Type Nuclear Membrane	-/+	?/+
Nuclear Pores and Membrane Bending Ability	-/+	-/+
Complex Endomembrane System	-/-	?/+

Fig 4.1: Comparison table of the Hypotheses. A "+" denotes that a theory proposes a credible explanation that is not precluded by most supporting primary data. A "-" indicates arguments that are not in line with reviewed research. A "?" denotes that a theory completely neglets to address a criteria. The symbol left of the slash refers to the original theory, while the symbol right of the slash refers to its modified version.

Nevertheless, the modified VE theory accounts for the origin of all 5 features whereas the modified syntrophy hypothesis can only explain the origin of 4 of them. We thus assess that the modified VE theory is superior to the modified syntrophy hypothesis.

The original VE theory deserves credit for being better able to accommodate modifications than the original syntrophy hypothesis as it better fits phylogenomic data indicating that most genes of the chimeric eukaryotic genome are of asgard ancestry, with most of the remainder being alpha-proteobacteria (*Dominant Contribution of Asgard Archaea to Eukaryogenesis | bioRxiv*). We assess that this is fundamentally due to the fact that when we compare the proposed contribution of other partners beyond asgard archaea and alpha-proteobacteria, the viral contribution in the VE theory is less extensive than the delta-proteobacterial in the Syntrophy hypothesis.

In the original VE theory, the proposed viral contribution in the VE theory involves only specific features of the nucleus like the protocoatomer system and linear chromosomes. Thus, when presented with contrary phylogenomic evidence, we can propose secondary loss as an explanation.

By contrast, in the original syntrophy hypothesis, the entire endomembrane system is of delta-proteobacteria ancestry. The possibility of secondary loss of the whole proto-endomembrane system midway through eukaryogenesis without outright killing the consortium is remote. The only possible fix would be to propose that the delta-proteobacteria host was eliminated entirely from the hypothesis. However, the origin of the eukaryotic endomembrane system from a non-alphaproteobacteria prokaryote like delta-proteobacteria is the defining feature of the syntrophy hypothesis, being maintained in all of its revisions (López-García and Moreira, “The Syntrophy Hypothesis for the Origin of Eukaryotes Revisited”; López-García and Moreira, “Selective Forces for the Origin of the Eukaryotic Nucleus”).

To test our hypothesis, that our modified VE theory is superior and likely can bear evidence to affirm it, we propose the following experiments/analysis:

1. A reconstruction of the genome of viruses of the *Tectillviricetes* family as they existed during eukaryogenesis (between 1.5 and 2.5 billion years ago(Richards et al.)). In particular, the presence of functional genes that confer protection against known Asgard antiviral defense mechanisms will help the modified VE theory account for the replacement of its nuclear membrane with one of bacterial type.
2. A phylogenomic study to ascertain the extent of viral contributions to the eukaryotic genome. This can involve comparing the reconstruction in 2. and the reconstruction of the LECA genome proposed by (Richards et al.) to identify orthologous relationships between LECA and viral genes. Evidence of viral ancestry of the vesicle coats or secondary loss of a viral derived procoatomer system favours the original and modified VE theory respectively.
3. An investigation of the mechanical viability of the mechanism for *de novo* endomembrane development we propose.

4. An investigation of the possibility of cell surface membrane replacement before endomembrane development. This investigation should examine both the phylogeny of genes encoding the bacterial lipid synthesis pathway and the mechanical viability of such an evolutionary transition.

Alternatively, the following parallel studies should target our modified syntrophy hypothesis.

1. A re-examination of the delta-proteobacterial contribution to the eukaryotic genome. As more ancient prokaryotic, archaeal and eukaryotic genomes are discovered and phylogenomic methods improve, our understanding of the relative contribution of each clade to LECA's genome must be reassessed continuously(Richards et al.). Indeed, a prominent delta-proteobacterial contribution to the eukaryotic endomembrane was identified in the past before being challenged by a more recent reassessment (Pittis and Gabaldón; *Dominant Contribution of Asgard Archaea to Eukaryogenesis | bioRxiv*). A new investigation may very well reaffirm the older interpretation, which would help our modified syntrophy hypothesis explain endomembrane evolution. Should this investigation also reveal delta-proteobacterial ancestry of vesicle coats, then our proposal of secondary loss of delta-proteobacteria nuclear pores is inferior to the original syntrophy hypothesis.
2. An examination of the evolutionary transitions responsible for the origin of Gram positive bacteria may offer significant insights on how to refine the modified syntrophy hypothesis's proposals for the origin of bacterial type nuclear envelopes.
3. An examination of which mechanisms for gene transfer from a host to an endosymbiont are viable. This would help refine the proposed modifications for the syntrophy hypothesis's explanation of gene transfer to the chimeric genome.

Works Cited

Bell, Philip J. L. "Eukaryogenesis: The Rise of an Emergent Superorganism." *Frontiers in Microbiology*, vol. 13, May 2022. *Frontiers*, https://doi.org/10.3389/fmicb.2022.858064.

Caforio, Antonella, et al. "Converting Escherichia Coli into an Archaebacterium with a Hybrid Heterochiral Membrane." *Proceedings of the National Academy of Sciences*, vol. 115, no. 14, Apr. 2018, pp. 3704–09. *pnas.org (Atypon)*, https://doi.org/10.1073/pnas.1721604115.

Dacks, Joel B., and Mark C. Field. "Evolution of the Eukaryotic Membrane-Trafficking System: Origin, Tempo and Mode." *Journal of Cell Science*, vol. 120, no. Pt 17, Sept. 2007, pp. 2977–85. *PubMed*, https://doi.org/10.1242/jcs.013250.

Dimitriu, Tatiana, et al. "Various Plasmid Strategies Limit the Effect of Bacterial Restriction–Modification Systems against Conjugation." *Nucleic Acids Research*, vol. 52, no. 21, Nov. 2024, pp. 12976–86. *Silverchair*, https://doi.org/10.1093/nar/gkae896.

Dominant Contribution of Asgard Archaea to Eukaryogenesis | bioRxiv. https://www.biorxiv.org/content/10.1101/2024.10.14.618318v1. Accessed 1 Dec. 2024.

Field, Mark C., and Michael P. Rout. "Coatomer in the Universe of Cellular Complexity." *Molecular Biology of the Cell*, vol. 33, no. 14, Nov. 2022, p. pe8. *PubMed Central*, https://doi.org/10.1091/mbc.E19-01-0012.

---. *Pore Timing: The Evolutionary Origins of the Nucleus and Nuclear Pore Complex*. 8:369, F1000Research, 3 Apr. 2019. *f1000research.com*, https://doi.org/10.12688/f1000research.16402.1.

Jékely, Gáspár. "Origin of Eukaryotic Endomembranes: A Critical Evaluation of Different Model Scenarios." *Advances in Experimental Medicine and Biology*, vol. 607, 2007, pp. 38–51. *PubMed*, https://doi.org/10.1007/978-0-387-74021-8_3.

Koonin, Eugene V. "The Origin of Introns and Their Role in Eukaryogenesis: A Compromise Solution to the Introns-Early versus Introns-Late Debate?" *Biology Direct*, vol. 1, Aug. 2006, p. 22. *PubMed Central*, https://doi.org/10.1186/1745-6150-1-22.

Lane, Nick. *Power, Sex, Suicide: Mitochondria and the Meaning of Life*. Oxford University Press, 2005.

Leão, Pedro, et al. "Asgard Archaea Defense Systems and Their Roles in the Origin of Eukaryotic Immunity." *Nature Communications*, vol. 15, no. 1, July 2024, p. 6386. *www.nature.com*, https://doi.org/10.1038/s41467-024-50195-2.

Liu, Yang, et al. "Virus-Encoded Histone Doublets Are Essential and Form Nucleosome-like Structures." *Cell*, vol. 184, no. 16, Aug. 2021, pp. 4237-4250.e19. *DOI.org (Crossref)*, https://doi.org/10.1016/j.cell.2021.06.032.

López-García, Purificación, and David Moreira. "Open Questions on the Origin of Eukaryotes." *Trends in Ecology & Evolution*, vol. 30, no. 11, Nov. 2015, pp. 697–708. *PubMed*, https://doi.org/10.1016/j.tree.2015.09.005.

---. "Selective Forces for the Origin of the Eukaryotic Nucleus." *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, vol. 28, no. 5, May 2006, pp. 525–33. *PubMed*, https://doi.org/10.1002/bies.20413.

---. “The Syntrophy Hypothesis for the Origin of Eukaryotes Revisited.” *Nature Microbiology*, vol. 5, no. 5, Apr. 2020, pp. 655–67. *DOI.org (Crossref)*, https://doi.org/10.1038/s41564-020-0710-4.

Makarova, Kira S., et al. “Unprecedented Diversity of Unique CRISPR-Cas-Related Systems and Cas1 Homologs in Asgard Archaea.” *The CRISPR Journal*, vol. 3, no. 3, June 2020, pp. 156–63. *PubMed Central*, https://doi.org/10.1089/crispr.2020.0012.

Martin, William F., et al. “Endosymbiotic Theories for Eukaryote Origin.” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 370, no. 1678, Sept. 2015, p. 20140330. *DOI.org (Crossref)*, https://doi.org/10.1098/rstb.2014.0330.

Martin, William, and Eugene V. Koonin. “Introns and the Origin of Nucleus–Cytosol Compartmentalization.” *Nature*, vol. 440, no. 7080, Mar. 2006, pp. 41–45. *DOI.org (Crossref)*, https://doi.org/10.1038/nature04531.

Martin, William, and Miklós Müller. “The Hydrogen Hypothesis for the First Eukaryote.” *Nature*, vol. 392, no. 6671, Mar. 1998, pp. 37–41. *DOI.org (Crossref)*, https://doi.org/10.1038/32096.

Martin, William, and Michael J. Russell. “On the Origins of Cells: A Hypothesis for the Evolutionary Transitions from Abiotic Geochemistry to Chemoautotrophic Prokaryotes, and from Prokaryotes to Nucleated Cells.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, vol. 358, no. 1429, Jan. 2003, pp. 59–83; discussion 83-85. *PubMed*, https://doi.org/10.1098/rstb.2002.1183.

Mendoza, Senén D., et al. “A Bacteriophage Nucleus-like Compartment Shields DNA from CRISPR Nucleases.” *Nature*, vol. 577, no. 7789, Jan. 2020, pp. 244–48. *DOI.org (Crossref)*, https://doi.org/10.1038/s41586-019-1786-y.

Pittis, Alexandros A., and Toni Gabaldón. “Late Acquisition of Mitochondria by a Host with Chimeric Prokaryotic Ancestry.” *Nature*, vol. 531, no. 7592, Mar. 2016, pp. 101–04. *PubMed Central*, https://doi.org/10.1038/nature16941.

Richards, Thomas A., et al. “Reconstructing the Last Common Ancestor of All Eukaryotes.” *PLOS Biology*, vol. 22, no. 11, Nov. 2024, p. e3002917. *PLoS Journals*, https://doi.org/10.1371/journal.pbio.3002917.

Wu, Fabai, et al. “Unique Mobile Elements and Scalable Gene Flow at the Prokaryote–Eukaryote Boundary Revealed by Circularized Asgard Archaea Genomes.” *Nature Microbiology*, vol. 7, no. 2, Feb. 2022, pp. 200–12. *www.nature.com*, https://doi.org/10.1038/s41564-021-01039-y.

Zachar, István, and Eörs Szathmáry. “Breath-Giving Cooperation: Critical Review of Origin of Mitochondria Hypotheses.” *Biology Direct*, vol. 12, no. 1, Aug. 2017, p. 19. *BioMed Central*, https://doi.org/10.1186/s13062-017-0190-5.

Effectiveness of Non-Surgical Rehabilitation for Rotator Cuff Tears in Athletes?
By Ryan Rahimi

Abstract

The rotator cuff comprises four key muscles around the shoulder: supraspinatus, infraspinatus, teres minor, and subscapularis. A rotator cuff tear occurs when any tendon connecting these muscles to the bones near the glenohumeral joint is damaged. Such injuries are especially common in sports involving repetitive rotating motions, like baseball, basketball, and lacrosse. For instance, in lacrosse, the repeated passing and shooting motions place the shoulder in excessive overhead external rotation, followed by powerful thrusts that create stress within the glenohumeral joint. These forces can ultimately result in rotator cuff tears. Recovery may involve either surgical or non-surgical approaches; however, several barriers exist to surgical intervention, such as high healthcare costs, existing medical conditions, and potential post-surgical complications. Due to these challenges, non-surgical methods are increasingly used to initiate rehabilitation and facilitate a return to play. Research highlights that non-surgical treatments like physical therapy, PRP (platelet-rich plasma) injections, NSAIDs (non-steroidal anti-inflammatory drugs), and cortisone injections are promising for rehabilitating rotator cuff tears. Physical therapy is essential for strengthening shoulder muscles, enhancing stability, and promoting tissue healing, while PRP injections leverage the body's growth factors for faster tissue repair. NSAIDs and cortisone injections help reduce inflammation and pain, allowing athletes to engage effectively in progressive rehab exercises. Implementing early treatment plans with these non-surgical interventions has shown positive effects on functional outcomes, pain reduction, and safe return-to-play rates, underscoring their critical role in rehabilitating athletes with rotator cuff tears.

Author Summary

Rotator cuff tears are a common injury among athletes, especially in sports like lacrosse that demand repetitive shoulder motions. These injuries can significantly impact performance and quality of life. Through this study, I explored non-surgical treatments for rotator cuff tears, focusing on their effectiveness in helping athletes return to play. Physical therapy, PRP injections, NSAIDs, and cortisone injections were examined for their ability to reduce pain, improve function, and promote recovery. Emerging treatments like extracorporeal shockwave therapy and ultrasound-guided procedures also showed promise in managing these injuries. My findings highlight the importance of early intervention and personalized rehabilitation plans for successful recovery. This research underscores the value of non-surgical approaches, offering athletes options to avoid surgery's lengthy recovery and potential complications. By understanding these alternatives, athletes, coaches, and healthcare providers can make informed decisions to support optimal recovery and long-term shoulder health.

Introduction

The rotator cuff encompasses four key muscles encircling the shoulder joint: the supraspinatus, responsible for internal rotation and arm elevation; the infraspinatus, facilitating external arm rotation within the shoulder socket; the teres minor, a compact muscle aiding arm rotation; and the subscapularis, governing arm abduction.[3] The integrity of this vital network is threatened when a tear develops in the tendon, which connects these muscles to the bones near the glenohumeral joint. Among athletes, repetitive motion is a common catalyst for this injury, often observed in athletes playing sports heavily predicated on upper arm movements.[1] For example, in lacrosse, players propel the shoulder into an excess of overhead external rotation, driven by the rhythmic passing and shooting motions. This can cause an injury to have profound disruption on individual player performance. In other sports as well, injuries may impact the overall team synergy as well as the individual's development and growth in the sport.[2] Of the various setbacks that athletes might encounter, rotator cuff tears emerge as a particularly frequent and debilitating concern, capable of significantly impeding an athlete's on-field prowess. The approach to addressing and recovering from rotator cuff tears can involve a combination of surgical and non-surgical techniques. Each avenue offers a unique array of advantages and considerations, carefully tailored to factors such as the tear's severity, the athlete's aspirations, and the preferred timeline for reentering the sporting arena. Illustrating the spectrum of interventions, surgical methods like arthroscopy emerge as a notable example. This minimally invasive procedure entails the insertion of a small camera through an incision, enabling surgeons to meticulously repair joint damage, as aptly described by the Mayo Clinic.[3] Alternatively, the option of open repair, in which an open surgical incision is made, comes into play, particularly in cases of substantial or intricate tears. In this traditional surgical approach, an arthroscope may be used, which is a tiny, tube-shaped instrument that is inserted into a joint. It consists of a system of lenses, a small video camera, and a light for viewing. The camera is connected to a monitoring system that allows the healthcare provider to view a joint through a very small incision. The arthroscope is very often used along with other tools that are inserted through another incision.[4] While these methods remain a staple in rotator cuff treatment, the realm of sports medicine has undergone a transformative evolution in recent years, marked by remarkable progress in the field of non-surgical interventions designed to rehabilitate athletes plagued by rotator cuff tears.[3] This progressive approach encompasses a fusion of evidence-based methodologies and highly individualized care protocols, culminating in a paradigm shift that redefines the trajectory of athletes' shoulder recovery. This article will review the effectiveness of these pioneering non-surgical interventions, illuminating their profound influence and pivotal role in empowering athletes to surmount the formidable challenges imposed by rotator cuff tears. Through a multifaceted lens, this article will uncover how these interventions have not only reshaped the course of injury management but have also become a beacon of hope, guiding athletes back to their peak performances.[3]

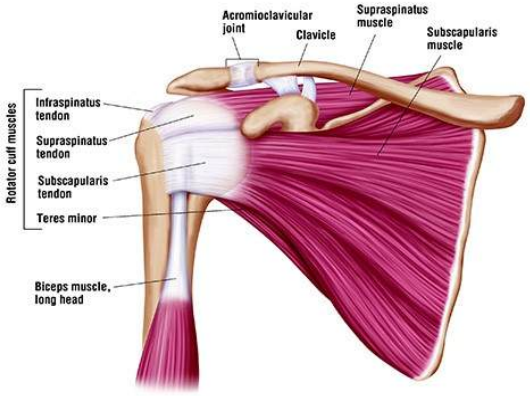


Figure 1: Figure of glenohumeral joint depicting the four rotator cuff muscles: supraspinatus, infraspinatus, teres minor, and subscapularis.

Methods

This research was conducted through a comprehensive literature review of scholarly articles and clinical studies. The terms "non-surgical rehabilitation for athletes" were searched in Google Scholar and PubMed to provide relevant sources for the review. Modifications of these search terms were made to broaden the scope of the search, and relevant articles' citations were also checked for possible inclusion. Articles between the years of April 2016 - September 2024 were included for possible analysis. Articles that focused on non-surgical rehabilitation approaches for rotator cuff tears in athletes were included. The review highlights a variety of non-surgical interventions and their reported efficacy in the context of athletic rehabilitation.

Body

The emerging body of research sheds light on the effectiveness of non-surgical interventions, particularly physical therapy, in the rehabilitation of rotator cuff tears.[4] Early initiation of treatment strategies has emerged as a critical factor in obtaining favorable and lasting outcomes. By engaging in physical therapy programs, individuals with rotator cuff injuries can experience significant improvements in the overall strength and stability of their shoulder joint, thereby fostering the process of tissue healing and recovery. PRP injection therapy stands as a treatment method that capitalizes on the concentrated bioactive agents found in platelets, delivering them directly to the site of injury to increase the intricate process of musculoskeletal tissue repair. Through this targeted approach, PRP injections hold promise in accelerating the healing process of rotator cuff tears, thereby potentially expediting the return to functional activities.[4]

Non-steroidal anti-inflammatory drugs (NSAIDs), rest, ice, and cortisone injections are integral components of managing rotator cuff tear symptoms; though not curative, these modalities play a critical role in the management of rotator cuff tears for patients pursuing both

surgical and non-surgical options. NSAIDs are commonly used for short-term pain relief and inflammation reduction, with a meta-analysis demonstrating that oral NSAIDs might be as effective as corticosteroid injections in alleviating short-term pain and improving function for those grappling with rotator cuff tendinopathy. Rest and activity modification are crucial, as patients are advised to avoid activities that exacerbate symptoms to allow damaged tissues to heal naturally.[5] Alternating between heat and ice therapy is recommended, where heat loosens tight muscles and joints, and ice reduces swelling and stiffness.[6] Cortisone injections play a vital role in managing the acute inflammation associated with rotator cuff tears, facilitating a smoother initiation of therapy for the patient. By significantly reducing inflammation, cortisone injections provide a window of relief that enables patients to engage in rehabilitation exercises with greater comfort. It is worth emphasizing the necessity of committing to the prescribed therapy regimen and exercises, even if the shoulder experiences relief following an injection. While cortisone effectively mitigates the immediate inflammation, the true healing and long-term prevention of recurrent rotator cuff tear symptoms are best achieved through consistent therapeutic exercises. These exercises not only contribute to the healing process but also strengthen the surrounding muscles and structures, thereby enhancing the overall stability of the shoulder joint.[8]

Patients might explore orthobiologics, a burgeoning field within regenerative medicine. Orthobiologics utilize the body's own resources to treat areas of inflammation, pain, or injury. Commonly used injections in this category include bone marrow aspirate concentrate injections (BMAC) as well as PRP injections. PRP injection therapy stands as a treatment method that capitalizes on the concentrated bioactive agents found in platelets, delivering them directly to the site of injury to increase the intricate process of musculoskeletal tissue repair. Through this targeted approach, PRP injections hold promise in accelerating the healing process of rotator cuff tears, thereby potentially expediting the return to functional activities.[4] Patients often have questions about these novel treatments, and these queries are addressed by experienced sports medicine providers with specialized training in this field.

Other emerging non-surgical methods include extracorporeal shockwave therapy (ESWT) and ultrasound-guided percutaneous irrigation of calcific tendinopathy (US-PICT). ESWT involves the application of high-energy shockwaves to the affected shoulder area, stimulating biological responses within the tissue. This therapy promotes the breakdown of calcific deposits, enhances blood flow to damaged tissue, and activates the body’s natural healing mechanisms, ultimately reducing pain and improving mobility. Studies have shown that ESWT is particularly effective in reducing pain in patients with chronic tendinopathies, including rotator cuff injuries, making it a valuable tool in non-surgical management for these conditions. In contrast, US-PICT is a minimally invasive procedure that uses ultrasound imaging to guide the removal of calcific deposits from the rotator cuff. This procedure involves inserting a small needle into the area of calcification and using fluid irrigation to dissolve and remove the deposits. By directly targeting the calcific buildup, US-PICT helps alleviate pain and restore function in the shoulder, offering quicker recovery with fewer risks compared to more invasive procedures. This technique shows

significant promise in treating rotator cuff calcific tendinopathy (RCCT), and its precision makes it an attractive option for patients seeking to avoid surgery.

Both ESWT and US-PICT provide non-surgical options that could be highly beneficial for patients suffering from RCCT, a condition often difficult to manage with traditional therapies alone. These treatments not only offer relief from pain but also improve long-term shoulder function, thereby expanding the range of non-surgical interventions available for rotator cuff injuries. Additionally, exercise therapy (ET), though not specifically tested for RCCT, has proven effective in treating rotator cuff tendinopathy in general. ET includes a structured regimen of strengthening and flexibility exercises aimed at improving shoulder mechanics, enhancing muscle coordination, and preventing further injury. By promoting gradual tissue repair and strengthening the surrounding muscles, ET can alleviate pain, restore range of motion, and improve functional outcomes, making it a valuable option within non-surgical rehabilitation programs.[16]

Understanding the progression of rotator cuff tears aids in devising personalized interventions that align with the specific nature and severity of the injury. Assessing the area of the tear, its underlying cause, and the duration since onset enables healthcare providers to formulate targeted treatment strategies. Differentiating between acute traumatic tears, chronic traumatic tears, or a blend of both is pivotal in determining the most suitable course of action. Moreover, recognizing the diverse array of symptoms exhibited by patients with rotator cuff tears necessitates a nuanced approach that encompasses not only the structural aspects but also the individual's lifestyle and functional requirements. By tailoring interventions to account for these multifaceted factors, healthcare professionals can optimize treatment outcomes and enhance the overall well-being of patients grappling with rotator cuff injuries.

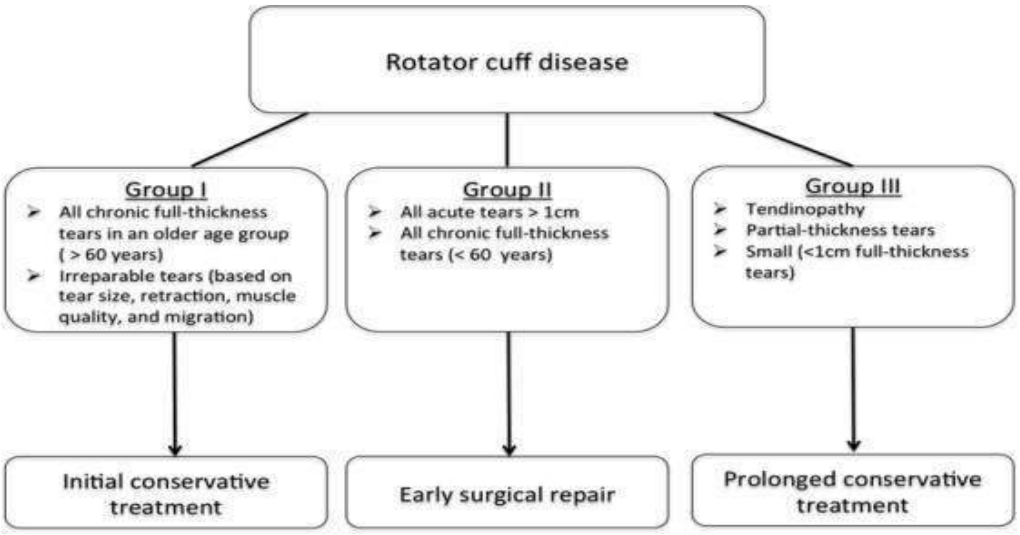


Figure 2: Treatment algorithm for pathology of the rotator cuff based on gradation of tears.

Discussion

This paper reviewed non-surgical interventions for managing rotator cuff tears. While a variety of methods are employed to help with pain and symptom management, such as NSAIDs, rest, ice, and cortisone injections, other non-surgical therapies attempt to heal rotator cuffs to the degree that surgery is no longer indicated. These therapies include PRP, targeted physical movements, and orthobiologics. However, while these interventions help manage pain and inflammation, they do not treat the underlying tear directly. Early treatment, patient compliance with therapy regimens, and individualized rehabilitation programs are critical for achieving optimal outcomes.

Despite the success of these non-surgical strategies, it is important to acknowledge that surgical intervention remains the mainstay of treatment for full-thickness rotator cuff tears. In scenarios where non-surgical interventions are ineffective, surgery is often necessary to restore full functionality and prevent long-term complications. Although non-surgical treatments are effective in relieving pain for many patients, it is important to acknowledge the possibility that the tear may enlarge without surgical repair. As a result, some patients may need to make long-term activity modifications to mitigate this risk. The decision to pursue surgical repair often relies on an individual's desire to regain full functionality and return to activities and sports without the limitations of pain, following a successful recovery period from surgery.[9] However, surgery comes with its own set of challenges and downsides. One of the primary concerns with surgery is the extended recovery time, which can last from six months to a year, during which athletes are unable to fully participate in their sport. This can be particularly problematic for athletes who need to time their recovery with competitive seasons or are under pressure to return to play quickly. In addition to the long recovery period, there is also the risk of surgical complications, such as infections, nerve damage, or stiffness in the shoulder joint, which could further delay recovery or negatively impact the athlete’s overall performance.

Moreover, while surgery can repair the structural damage, it does not always guarantee a return to pre-injury levels of function, especially for older athletes or those with chronic tears. There is also the risk of re-tearing the rotator cuff, which can lead to a cycle of repeated surgeries and prolonged recovery periods. Financial limitations can also pose a barrier, as surgery is often more costly than conservative treatment options and may not be covered fully by insurance, especially for athletes in non-professional settings.

The context of athletes is particularly relevant, as they may face barriers to surgery, such as the timing of their athletic seasons, the desire to avoid long recovery periods, and financial limitations. In such cases, non-surgical interventions become invaluable tools for helping athletes return to play in a timely manner. When applying this knowledge to different sports, it becomes clear why certain athletes must be aware of these methods. In sports like tennis, baseball, and swimming, which all involve overhead motions, understanding when to pursue non-surgical versus surgical interventions can significantly impact an athlete’s recovery timeline and performance. The nature of the sport, the severity of the tear, and the athlete's professional goals all play a role in determining the best course of action.

The treatment algorithm for triaging tears—determining when an athlete should pursue non-surgical options versus surgical repair—should be based on the severity of the tear and the athlete's recovery goals. Athletes with minor or partial tears may benefit significantly from non-surgical approaches, while those with full-thickness tears may require surgery for a full recovery (Figure 1). However, when considering surgery, both athletes and healthcare providers must weigh the potential for a full structural repair against the time off from sports, the risks of complications, and the long-term outcomes, which may not always result in complete restoration of shoulder function.

This review highlights the effectiveness of non-surgical interventions in rehabilitating rotator cuff tears, particularly for athletes. The strengths of this review lie in its comprehensive exploration of different non-surgical methods and their practical implications for athletes. Future research should focus on comparing outcomes in different sports and evaluating long-term effects of non-surgical rehabilitation versus surgical interventions. As advances in regenerative medicine, like PRP and orthobiologics, continue to evolve, future studies should assess the long-term efficacy of these treatments in sports-related rotator cuff injuries. By exploring these diverse non-surgical options, this review contributes to the growing body of evidence supporting the effectiveness of conservative treatment approaches. This knowledge is essential for athletes, coaches, and healthcare providers who are navigating treatment decisions for rotator cuff tears in the demanding world of sports.

Conclusion

Rotator cuff tears are a common issue among athletes, arising from the intense demands placed on the shoulder joints. Repetitive throwing motions, sudden directional changes, and physical collisions during gameplay place significant stress on these joints. Effective rehabilitation is critical not only for restoring functionality but also for preventing re-injury, enhancing athletic performance, and promoting psychological well-being, all while ensuring the long-term health of the shoulder. Although gaps in research persist, existing evidence highlights the effectiveness of non-surgical interventions in successfully rehabilitating athletes with rotator cuff tears. This underscores the importance of developing tailored approaches to address the specific needs and demands of athletes in various sports.

Works Cited

Jack, R. A., et al. "Full-Thickness Rotator Cuff Tears in the Throwing Athlete." Operative Techniques in Sports Medicine, vol. 29, no. 1, 2021, p. 150800. https://doi.org/10.1016/j.otsm.2021.150800.

Hurley, O. A. "Impact of Player Injuries on Teams' Mental States, and Subsequent Performances, at the Rugby World Cup 2015." Frontiers in Psychology, vol. 7, 2016, p. 807. https://doi.org/10.3389/fpsyg.2016.00807.

Cleveland Clinic. "What Is the Anatomy of the Rotator Cuff?" Cleveland Clinic, n.d. Accessed 5 Jan. 2024, https://my.clevelandclinic.org/health/body/rotator-cuff.

Johns Hopkins Medicine. "Rotator Cuff Repair." Johns Hopkins Medicine, 8 Aug. 2021, https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/rotator-cuff-repair.

Shoulder Surgeon in Seattle. "Rotator Cuff Tear - Surgical and Non-Surgical Treatments." Shoulder Surgeon in Seattle, 21 June 2021, https://seattleshoulderdoc.com/rotator-cuff-tear/.

Weiss, L. J., et al. "Management of Rotator Cuff Injuries in the Elite Athlete." Current Reviews in Musculoskeletal Medicine, vol. 11, no. 1, 2018, pp. 1–8. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5825345/.

Johnston, P. "Rotator Cuff Tears (Non-Surgical Treatment)." Dr. Peter Johnston, n.d. Accessed 23 Aug. 2023, https://www.marylandshoulderelbow.com/nonsurgical-treatment-of-rotator-cuff-tears.

Cluett, J. "8 Ways to Treat Your Rotator Cuff Tear Without Surgery." Verywell Health, 1 May 2022, https://www.verywellhealth.com/non-surgical-treatments-for-rotator-cuff-tears-2549784.

Tolbert Center for Rehabilitation and Wellness. "PRP Injections: The Best Non-Surgical Treatment for Rotator Cuff Tears?" Tolbert Center for Rehabilitation and Wellness, 5 Mar. 2020, https://drglennatolbert.com/2018/01/prp-injections-best-non-surgical-treatment-rotator-cuff-tears/.

Summa Health. "Torn Rotator Cuff? Try These 4 Non-Surgical Treatments That Really Work." Summa Health, n.d., https://www.summahealth.org/flourish/entries/2023/07/torn-rotator-cuff-try-these-4-non-surgical-treatments-that-really-work.

University of Washington Orthopedics and Sports Medicine. "Rotator Cuff Tear: Non-Operative Treatment." University of Washington Orthopedics and Sports Medicine, n.d., https://orthop.washington.edu/patient-care/shoulder/rotator-cuff-tear-non-operative-treatment.html.

Williams, C. "Are There Nonsurgical Treatment Options for a Rotator Cuff Tear?" Atlanta, GA Orthopedic Specialist, 31 Dec. 2021, https://ioaregenerative.com/blog/are-there-nonsurgical-treatment-options-for-a-rotator-cuff-tear.

Edwards, P., et al. "Exercise Rehabilitation in the Non-Operative Management of Rotator Cuff Tears: A Review of the Literature." International Journal of Sports Physical Therapy, vol. 11, no. 2, Apr. 2016, pp. 279–307. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4827371/.

Bieber, E. "Rotator Cuff Disorders: The Facts." Ortho Bethesda, 30 Nov. 2022, https://www.orthobethesda.com/blog/rotator-cuff-disorders-the-facts/.

Mayo Clinic. "Arthroscopy." Mayo Clinic, 19 Aug. 2022, https://www.mayoclinic.org/tests-procedures/arthroscopy/about/pac-20392974.

Caballero, I., et al. "Effectiveness of Non-Surgical Management in Rotator Cuff Calcific Tendinopathy (The Effect Trial): Protocol for a Randomised Clinical Trial." BMJ Open, vol. 14, no. 1, 2024, e074949. https://doi.org/10.1136/bmjopen-2023-074949.

Edwards, P., et al. "Exercise Rehabilitation in the Non-Operative Management of Rotator Cuff Tears: A Review of the Literature." International Journal of Sports Physical Therapy, 2016, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4827371/.

Bieber, E. "Rotator Cuff Disorders: The Facts." Ortho Bethesda, 30 Nov. 2022, https://www.orthobethesda.com/blog/rotator-cuff-disorders-the-facts/.

Using Machine Learning to Predict ChatGPT's Mathematical Problem-Solving Capabilities and Identify Factors Affecting Its Performance

By Yadnya Patil

Abstract

"Artificial Intelligence is the automation of cognition."(Abbass, 2021). Understanding how they process the problem details is crucial for improving the model's problem-solving performance and refining its mathematical reasoning capabilities. This research aims to predict and analyze the factors influencing ChatGPT's ability to solve mathematical problems using a dataset of 10,000 diverse math questions covering topics like algebra, calculus, and geometry, ranging from middle school to college-level content. It includes word problems, equations, and other problem types. By engineering features such as problem difficulty, solution complexity, and question tone, the study incorporates these factors into a classification model to identify the key elements that impact ChatGPT's performance. Machine learning techniques, particularly Random Forest classifiers, were trained to predict chatGPT's performance and assess factors using feature importance analysis. The model achieved an accuracy of 93%, demonstrating a strong predictive capability in determining the ChatGPT's performance. Feature importance analysis revealed that the difficulty level of the question, the complexity of the solution, and the model's performance on the previous question (where a wrong answer increased the likelihood of subsequent errors) played significant roles in influencing performance. These findings provide valuable insights for improving ChatGPT's mathematical reasoning skills and suggest ways for the user to refine their approach to solving complex problems.

Introduction

Artificial Intelligence (AI) has made significant advances, particularly in Natural Language Processing, with Large Language Models (LLMs) like ChatGPT emerging as powerful tools for various tasks. One such area is mathematical problem-solving, where LLMs can provide solutions to problems ranging from basic arithmetic to complex equations. Friede et al. (2023) examined ChatGPT's inconsistent performance in advanced mathematics (noting that specialized models outperformed it) but recognized its value as an assistant for proficient users. Poola and Božić (2023) proposed a framework that integrates machine learning with human intuition to aid mathematicians and uncover patterns and connections in mathematics, though challenges like large dataset requirements remain. Overall, these studies showcased AI's promise in education and mathematics.

However, despite their potential, AI's performance in mathematics remains inconsistent, especially when dealing with problems of varying difficulty and complexity (Hong et al., 2020). Mathematical problem-solving requires logical reasoning, pattern recognition, and precision (Patel et al. 2021). Previous studies (OpenAI , 2024) have emphasized that LLMs often struggle with multi-step or abstract reasoning tasks. Furthermore, Zong et al. (2023) identified that GPT models perform well in classification problems, converting descriptions to equations (79%

success), and generating similar problems. There remains a need to understand the factors that influence and methods to optimize LLM performance in solving complex mathematical problems.

To my knowledge, the current work is the first to systematically evaluate the importance of factors like problem difficulty, problem complexity, and tone of the prompt on LLM performance. Unlike prior research, this study focuses on using machine learning and feature importance analysis to predict and analyze the mathematical reasoning capabilities of LLMs, offering new insights into improving AI for educational contexts.

Dataset Compilation

The dataset for this study, described in Saxton et al. (2019) (arXiv, 2019), consists of 10,000 carefully selected mathematical problems (Figure 1) spanning diverse subject areas. Primarily numerical, the problems include a mix of arithmetic, algebra, geometry, and calculus, offering a comprehensive representation of mathematical tasks. Each problem was paired with its correct solution to evaluate the model's accuracy in providing correct answers.

```
Question: Solve -42*r + 27*c = -1167 and 130*r + 4*c = 372 for r.
Answer: 4
Question: Calculate -841880142.544 + 411127.
Answer: -841469015.544
Question:  Let x(g) = 9*g + 1.  Let q(c) = 2*c + 1.  Let f(i) = 3*i -
39.  Let w(j) = q(x(j)).  Calculate f(w(a)).
Answer: 54*a - 30
Question: Let e(l) = l - 6.  Is 2 a factor of both e(9) and 2?
Answer: False
Question: Let u(n) = -n**3 - n**2.  Let e(c) = -2*c**3 + c.  Let l(j)
= -118*e(j) + 54*u(j).  What is the derivative of l(a)?
Answer: 546*a**2 - 108*a - 118
Question:  Three letters picked without replacement from qqqkkklkqkkk.
Give prob of sequence qql.
Answer: 1/110
```

Figure 1. Examples from the Dataset

To gather responses, each question in the dataset was passed through a Python program, which queried ChatGPT for an answer. The prompt template is as follows:
"Solve this problem: {actual problem}"

ChatGPT was given 30 seconds to respond to each question, which is sufficient for generating a full response. During this process, additional features were extracted from ChatGPT's responses. These extracted features are used alongside the dataset's inherent characteristics to evaluate ChatGPT's performance comprehensively.

Engineered Features

Though the dataset did not have predefined features, several relevant features were identified that could potentially influence ChatGPT's performance. According to Wardat Yousef's paper, LLMs face several challenges in solving math problems, such as difficulties in

handling multi-step reasoning, adapting to different tones in problem phrasing, and maintaining accuracy across diverse problem domains. To address these challenges, we categorized the features based on both the quantitative and qualitative aspects of the problems, considering the specific characteristics of each task.

1. **Difficulty Level**
   Problems were categorized as low, medium, or high difficulty based on how ChatGPT assessed the complexity of the concepts involved. The model evaluated each problem and assigned it a difficulty level accordingly.
   Prompt: *"Please rate the difficulty of the following problem as low, medium or high"*
2. **Number of Steps in Solution:**
   The number of steps was determined by ChatGPT's breakdown of the problem, though it may vary as different solvers may approach the problem differently.
   Prompt: *"How many steps did it take you to solve this question"*
3. **Question Tone:**
   The tone of the question, whether formal or conversational, was analyzed to understand its effect on ChatGPT's performance. This factor helped examine how phrasing influenced the accuracy of the model's responses.
   Prompt: *"Do you think that this question's tone was friendly, formal or instructional?"*
4. **Error Propagation:**
   After all the questions went through the Python program and ChatGPT generated all the answers, it was then provided with the correct solutions and prompted to compare them to the generated answers to assess response accuracy. The correct solutions were provided to ChatGPT sequentially, each followed with the following prompt: *"How many times were two consecutive problems answered incorrectly?"* For each sample, this feature captures the current performance history of the LLM.
5. **Source of Problem:**
   Problems were already categorized based on their domain (e.g., algebra, calculus, geometry) in the database. Different domains often involve distinct problem-solving strategies, and some may be more difficult due to their mathematical nature or the computational techniques required.

Data Preprocessing

To prepare the dataset for model training and evaluation, several preprocessing steps were applied:

1. **Text Normalization**: Problem statements were standardized to remove any inconsistencies in wording and ensure clarity. Ambiguous or unclear statements were identified when ChatGPT had difficulty providing correct or clear answers. Since

ChatGPT works best with Python code, the problem statements were converted (by ChatGPT) into Python scripts to ensure clarity.

2. **Mathematical Expression Standardization**: All mathematical expressions were converted into a uniform format that could be easily processed by the model. This included ensuring that symbols and operations were correctly represented.

The dataset was split into two main subsets: a training set and a testing set. 80% of the data (8,000 samples) was used for training, while the remaining 20% (2,000 samples) was reserved for testing. This methodology, which aligns with Fischer et al. (2023) and Cattan et al. (2021), ensures that the model is trained on a sufficiently large portion of the data while retaining an adequate test set for evaluating generalization performance.

Machine Learning Models

Logistic Regression and the Random Forest Classifier were trained on the binary classification task of predicting whether chatGPT would correctly answer a question based on the features extracted from that question. In case of imbalance in the label class distribution, the dataset's class imbalance was addressed using the SMOTE oversampling techniques. The models were evaluated using accuracy, precision, recall, and F1-score. Feature importance analysis was employed to uncover key drivers of ChatGPT's success.

Results and Analysis

1. **Model Performance**

   The Random Forest Classifier outperformed logistic regression, achieving 93% accuracy compared to 85%. The F1-score for the random forest model was 0.92, indicating robust predictive capability.

	Precision	Recall	F1-score	Support
Class 0	0.94	0.92	0.93	987
Class 1	0.92	0.93	0.92	1013
Accuracy			0.93	2000
Macro Avg	0.93	0.92	0.93	2000
Weighted Avg	0.93	0.93	0.93	2000

Table 1. The classification report

2. **Feature Importance**

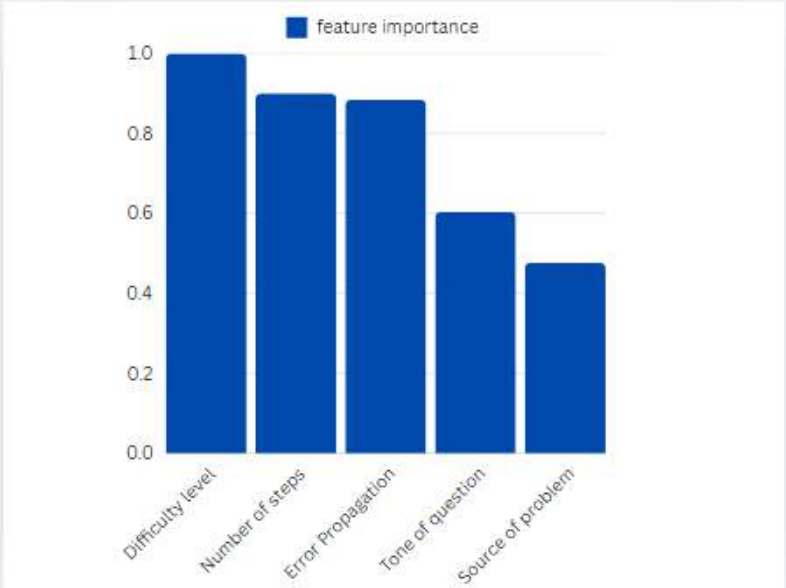


Figure 2. Feature Importance Chart

3. **Error Analysis**

Common error patterns included: 1) misinterpreting multi-step problems, and 2) difficulty with implicit reasoning and ambiguous questions.
Through manual experimentation with ChatGPT, errors and limitations were identified across different methods of input, i.e. the same questions were tested in three different ways to observe variations in accuracy and problem-solving approach.

a) Input a single problem at a time

In this scenario, a single question was prompted to ChatGPT at a time, and ChatGPT accurately interpreted mathematical notations and provided clear, step-by-step explanations that enhanced understanding. The final answers were distinctly boxed and bolded for easy identification.

b) Input two problems at a time

In this scenario, two random questions were prompted to ChatGPT at a time. It was observed that ChatGPT would often (but not always) only answer one of the questions and skip the other one, citing limited data processing capacity. This shows ChatGPT's limitations in handling multiple input questions at once.

c) Input the entire file of questions at once

There is a character limit for a single input, so when the file is too large, ChatGPT struggles to process it. To address this, the file was split into two parts. However, even with the

provided cell numbers, ChatGPT was unable to extract the questions from the file. The file was originally a Google Sheet.

The experiments demonstrated that ChatGPT struggles to process multiple input questions simultaneously and that asking one question at a time leads to fewer errors than presenting several questions together. This suggests that the optimal approach is to prompt ChatGPT with a single question at a time.

Discussion and Future Work

Feature importance analysis revealed the following, listed in order from most to least important:

1. **Difficulty level**: the most significant predictor of performance. Accuracy decreased as difficulty increased.
2. **Number of steps**: Problems requiring multiple steps to solve were classified as high complexity. These types of problems often demand more reasoning and deeper understanding, as they involve a series of logical steps to reach the correct answer.
3. **Error Propagation**: This feature indicates that a wrong answer increases the likelihood of a subsequent wrong answer. It demonstrates that errors tend to be repeated, while correct answers often lead to more correct responses. Put another way, if ChatGPT keeps getting questions wrong, then it's not a good idea to keep asking more questions right away. Instead, starting a new chat or session might help, as it can give ChatGPT a fresh start and improve the chances of getting the right answers next time. This idea should be considered when thinking about how past mistakes influence future responses.
4. **Tone of question**: When the user adopts a polite or friendly tone, the likelihood of receiving a correct answer increases. Conversely, if the user is angry or scolds ChatGPT for making mistakes, the model is more likely to provide an incorrect answer.
5. **Source of problem**: This feature had minimal impact on the results.

In future work, the plan is to categorize the data by assigning population percentages to each group, which will clarify how errors are distributed across different types of math problems. By analyzing these breakdowns, we can identify trends or patterns in the mistakes, potentially revealing areas for system improvement. This deeper analysis will help refine problem-solving methods and lead to more accurate results.

Conclusion

In this research, the mathematical problem-solving capabilities of ChatGPT were explored, and factors influencing its performance were identified using machine learning techniques. AI models like ChatGPT have shown great potential, particularly in natural language processing, but their ability to solve complex mathematical problems remains inconsistent, especially when faced with varying levels of difficulty and the need for multi-step reasoning. The need for a deeper understanding of how these models process mathematical tasks is crucial for

improving their performance in educational and problem-solving contexts.

By compiling a dataset of 10,000 diverse math problems from subjects such as algebra, calculus, and geometry, this study aimed to identify the key features affecting ChatGPT's ability to solve these problems accurately. Features such as problem difficulty, solution complexity, number of steps involved, tone of the prompt, and error propagation were extracted to understand their influence on ChatGPT's performance. The analysis utilized machine learning models, including Random Forest and Logistic Regression, to predict the model's ability to solve the problems, with the Random Forest model achieving 93% accuracy.

The results indicated that the most influential factors on ChatGPT's performance were: 1) the difficulty level of the problem, 2) the number of steps required to reach the solution, and 3) the propagation of previous errors. These findings suggest that ChatGPT's accuracy tends to decrease with increased problem difficulty and complexity. Error propagation also played a significant role, where incorrect answers were more likely to be followed by additional errors. The study found that problem tone had a lesser but still notable effect, with a more friendly or formal tone leading to better performance.

In terms of improving ChatGPT's mathematical problem-solving abilities, the findings emphasize the importance of clear, well-defined questions, presented one at a time, to enhance accuracy. Ambiguous or overly complex inputs tended to lower the model's performance. Additionally, giving feedback on mistakes and managing error propagation by refreshing the model's context may help improve its responses. Future work will delve deeper into categorizing different problem types and understanding error distributions for each of them separately to refine ChatGPT's mathematical reasoning and further enhance its problem-solving capabilities.

Work Cited

Abbass, Hussein. "What Is Artificial Intelligence?" IEEE Transactions on Artificial Intelligence, vol. 2, no. 2, 2021, pp. 94-95, https://ieeexplore.ieee.org/abstract/document/9523786.

Frieder, Simon, et al. "Mathematical Capabilities of ChatGPT." Advances in Neural Information Processing Systems, vol. 36, 2024, proceedings.neurips.cc/paper_files/paper/2023/hash/58168e8a92994655d6da3939e7cc0918-Abstract-Datasets_and_Benchmarks.html.

Hong, Yining, et al. "Learning by Fixing: Solving Math Word Problems with Weak Supervision." AAAI, 2020, ojs.aaai.org/index.php/AAAI/article/view/16629.

OpenAI. "OpenAI Announces a New AI Model, Code-Named Strawberry, That Solves Difficult Problems Step by Step." Wired, 2024, www.wired.com/story/openai-o1-strawberry-problem-reasoning.

Patel, Arkil, et al. "Are NLP Models Really Able to Solve Simple Math Word Problems?" arXiv, 2021, arxiv.org/pdf/2103.07191.

Poola, I., & Božić, V. (2023). Guiding AI with human intuition for solving mathematical problems in ChatGPT. https://www.researchgate.net/profile/Indrasen-Poola/publication/373447147_Guiding_AI_with_human_intuition_for_solving_mathematical_problems_in_Chat_GPT/links/64ece15eee5ca306008c1f19/Guiding-AI-with-human-intuition-for-solving-mathematical-problems-in-Chat-GPT.pdf

Saxton, David, et al. "Analytical Datasets for Mathematical Problem Solving." arXiv, 2019, arxiv.org/abs/1904.01557.

Wardat, Yousef, et al. "ChatGPT: A Revolutionary Tool for Teaching and Learning Mathematics." European Journal of Mathematics and Science Teaching, 2023, ejmste.com/article/chatgpt-a-revolutionary-tool-for-teaching-and-learning-mathematics-13272?trk=article-ssr-frontend-pulse_x-social-details_comments-action_comment-text.

Zong, Mingyu, et al. "Solving Math Word Problems Concerning Systems of Equations with GPT Models." ScienceDirect, 2023, www.sciencedirect.com/science/article/pii/S2666827023000592.

Microplastic Management Policies in Personal Care and Cosmetics Products: An Overview of Challenges and Prospects in China with a Comparative Analysis of the United States

By Zilin Xiang

Abstract

Microplastics in personal care and cosmetics products (PCCPs) have garnered significant attention in research concerning their health and ecological consequences. However, studies on policies aimed at regulating their sources remain under-explored, particularly in the context of China. This paper adopts a comparative approach to analyze policies addressing microplastics in PCCPs between the United States and China. The United States, having enacted the Microbead-Free Waters Act years before similar initiatives in China, offers a more developed framework that could serve as a reference for China's relatively recent work in environmental cosmetic regulations. I structured results of policy content analysis according to four key dimensions: goals, targets, instruments, and agents. Findings show that while efforts have been made to ban microbeads in PCCPs, their effectiveness is undermined by several challenges: ongoing debates about definitions for microplastic and the emergence of nanoplastics, unreliable detection methodologies, and a lack of unified understanding and awareness among key stakeholders—scientists, policymakers, industry players, and the public. The findings provide valuable insights for scholars interested in emerging environmental pollutants, legislators, the cosmetics industry, and eco-conscious consumers.

Introduction

Microbeads are plastic particles less than 5mm in diameter used in personal care and cosmetics products for functions such as exfoliation, scrubbing, or cleaning (United Nations Environment Programme [UNEP]). Microplastics accumulating in the food chain and/or absorbing toxicants as they travel through the environment are suspected to pose concerns for human health (Eriksen et al.).

On a global scale, the sum of annual emissions of microplastics from PCCPs in Europe, the United States, and China was 3843 tons in 2015, and an estimated 3.0×10^5 tons of PCCP-derived microplastics is believed to have accumulated in the environment in the last five decades (Sun et al.). Imagine a line of shipping containers filled only with microplastics: the annual emissions in 2015 would fill 137 containers, and the 300,000 tons accumulated over five decades would require more than 10,700 containers. They mostly are directly released into the aquatic environment with domestic sewage and escape from wastewater treatment plants (WWTPs), and have been reported to contaminate the aquatic environment, and are sufficiently small to be readily ingested by aquatic organisms (Cheung and Fok). The goal, then, of a national/international policy banning microbeads is to prevent the introduction of an otherwise avoidable choice of substance into the water supply (Strifling). This paper examines the microbead management policies in the United States and China, two of the largest PCCP markets globally, highlighting legal loopholes and potential areas for improvement in the current policies

of both countries. Studying how different governments handle the same problem can help identify why their approaches vary, which can lead to new ideas and theories (Gupta). Findings show that while efforts have been made to ban microbeads in PCCPs, their effectiveness is undermined by several challenges. These include ongoing debates about definitions of microplastics and the emergence of even smaller nanoplastics, unreliable detection methodologies, and a lack of unified understanding and awareness among key stakeholders—scientists, policymakers, industry players, and the public. The findings provide valuable insights for scholars interested in emerging environmental pollutants, as well as for policy analysts, legislators, the cosmetics industry, and eco-conscious consumers.

Literature Review

In recent decades, there has been a significant proliferation of monitoring studies on microplastics (Zitko and Hanlon; Gregory; Derraik; Thompson et al.; Fendall and Sewall; Arthur, Baker, and Bamford; Leslie, Moester, de Kreuk, and Vethaak; Leslie). Many countries and regions have acknowledged the need to reduce plastic microbeads in PCCPs, and socially responsible manufacturers have started phasing them out. However, challenges related to cost, detection methods, and regulatory measures indicate that considerable progress is still required (Chen, Li, & Xu; Zhang et al.; Sorensen et al.).

In determining appropriate intervention methods and distribution of responsibilities among stakeholders, the UNEP emphasized prevention over remediation, given the vast scale of the issue and the ecological risks of cleanup efforts (UNEP). Scholars widely advocate for extended producer responsibility, emphasizing that microplastics in PCCPs are an "intentionally added" pollutant requiring proactive industry action (Galafassi et al.; Andrady; Foster & Environment; Sherrington et al.; Wang et al.).

Nevertheless, microplastic regulation in PCCPs, like other environmental policies, face institutional challenges that limit its effectiveness. It is important to compare the practices and solutions in different countries, producing an evaluation for factors of successful—or unsuccessful—intervention. As we see later in the comparison, China and the US had to naviagate through priorities and instruments for tackling the following problems: first, the technical complexity of decisions in environmental policies are only understood by scientists and to some degree by politicians, but its success requires participation by all stakeholders (Gawlik et al.). Second, limited resources mean that not all problems can be tackled simultaneously, yet current approaches often address issues based on private interest or public nuisance rather than their ecological consequence (Fiorino). Third, scope and feasibility have a negative correlation. Critics argue that the Microbead-Free Waters Act (MFWA), for instance, was too narrow, excluding non-cosmetic microbeads or those used for purposes other than exfoliation (Graney). Nonetheless, the Act's limited scope made it politically feasible (Strifling). Four, the potential environmental consequences are often glanced over compared to direct and immediate symptoms of harm on the human body. It was observed that many cosmetic companies are rarely compelled by law to consider the environmental implications of their raw materials, as most of the

regulations are focused on the safety of the products for direct human use (Zhou et al.; Leslie et al.).

Effective environmental policies should maintain consistent standards worldwide, as their adverse effects on human health and the environment are generally the same. Nevertheless, there are ongoing debates in policymaking, namely the definition and scope of microplastic in PCCPs, justifying the need for a comparison and inquiry into the political, economic considerations behind such decisions. The 2015 UN plastic in cosmetics report defines microbeads as solid particles ranging from 1 to 1000 μm, as distinguished from microspheres which is similar in size but restricted to a spherical shape, and microcapsules, which is only around 1-2 μm, whereas the current policies in China and the United States adopt a broader limit of 5 mm (equivalent to 5000 μm; see "Detailed Standards for the Prohibition and Restriction of Related Plastic Products [2020 Edition]" and the MFWA, respectively). Polymers such as polyethylene (PE), polypropylene (PP), and polymethyl methacrylate (PMMA) are indisputably banned. However, ambiguities arise with "skeptical microplastics," including poloxamers, PEGs, and certain polyquaternium, which are sometimes excluded from regulations (Thompson et al.; Arthur et al.).

Although the impacts of microplastics are well documented, there is a noticeable gap in policy studies focused on mitigating microplastics in PCCPs, particularly in China. Within the few studies relevant to this topic, the Microbead Free Waters Act (MFWA) was mentioned alongside microplastic regulation of other countries to illustrate the need for control of microbeads in China (Zou & Li; Shui; Zhang et al.; Huang et al.; Zheng & Chen).Other US policies related to cosmetic supervision and labeling such as Modernization of Cosmetics Regulation Act (MoCRA) has also received some attention (Su et al., Wang & Zhu). Nevertheless, none have specifically compared the content of policy documents related to microplastic regulation in China and the US. In the US, one study estimated that MFWA will prevent 8 trillion MP microbeads from entering the nation's waters (Rochman et al.). Another set of results show that median concentrations of irregular microbeads, composed of polyethylene plastic declined by up to 86% in WWTP effluents after the bans, while those of spherical microbeads, predominantly synthetic/polyethylene wax did not differ, since, as categorized as non-plastic legally, they were not regulated (despite questionable biodegradability) (Suaria, et al.). Amounts of irregular microbeads declined relative to spherical microbeads in Lake Ontario, indicating that product changes may be influencing observations in lake waters. The results suggest that the United States restrictions effectively and rapidly reduced plastic microbeads entering waters via WWTPs (Akhbarizadeh et al.). In China, however, multiple tests show limited effectiveness of the bans. A research on rinse-off cosmetics and toothpaste products available on the market from 2020 to 2021 showed a detection rate of 35% for plastic microbeads. The primary detected plastic components were polyethylene and polyvinyl chloride, with some product labels explicitly indicating the presence of polyethylene scrub particles. When sorting the detected products by production date, a transition was observed from single plastic microbeads to a mixture of plastic microbeads and plant-based particles (Zhang et al.). This trend reflects the gradual replacement of plastic microbeads, but at a significantly slower pace than

other major economies, and thus justify the need for this study to evaluate legislation progress and limitations in China. Another research detected plastic microbeads in 6 out of 43 batches of daily chemical products tested in 2021, including scrubs, cleansers, gels, facial washes, and toothpaste. All detected products were manufactured before 2017. In 2022, the NMPA repeated the tests on rinse-off cosmetics and toothpaste products, finding no plastic microbeads in any samples (Huang et al.). Surveys indicate that even after the release of the *2019 Industrial Structure Adjustment Guidance Catalogue* and the *Opinions on Further Strengthening Plastic Pollution Control*, over 70% of respondents were unaware of the presence of plastic microbeads in skincare products or their environmental harm. Moreover, studies conducted after the 2020 bans found that sampled exfoliating gels and scrubs still contained 1-5% plastic microbeads, demonstrating limited regulatory effectiveness on manufacturers.

The gradual phase-out of plastic microbeads in PCCPs has become an inevitable global trend. Based on this preliminary assessment, recommendations were made to support current and future policy on national and regional plastic management strategies. Reviewing relevant literature on the topic shows that there are relatively few studies which offer a comprehensive comparison of all relevant policies related to microplastics in PCCPs in China and the US, especially given that the field is rapidly evolving, many studies have become incomplete or outdated. Recommendations to mitigate microplastics in PCCPs include improved practices for: (1) cohesion between different environmental laws and departments; (2) education, outreach and public awareness; (3) market-based practices to encourage businesses; and (4) increased monitoring and further research.

Methodology

The methodology of the paper is content analysis based on relevant regulations found in both countries. The research process consists of two stages: policy selection and content analysis. The policy documents selected for the policy content analysis have been listed in Table 1. At the time of carrying out this research and drafting the manuscript (October 2024 – December 2024), these constituted the most relevant policy documents and were hence included in the study. This paper manually sorts the content of policy documents, because current definitions and emphasis on microplastic cannot be clearly identified through quantitative analysis with software tools. The selection of policy texts in this research abided by the following principles:

- **Relevance**: Policy documents related to “microbeads in PCCPs” and policy documents with similar keywords (microplastics, nanoplastics, microspheres, nanospheres, plastic particulates, microcapsules, nanocapsules, cosmetics, skincare).
- **Authority**: The texts of laws, regulations and policies are all issued by the central government of the two countries, and priority given to specific departments responsible for environmental pollution.

- **Completeness**: The time limit for selecting the texts is from 2018 to 2024, and statistics are based on the environmental policy texts involved during this period, therefore examining the whole process to the emergence of such laws in China and the US.

Table 1. Summary of Policies Relating to "Microbeads in Cosmetics" in China

Year	Title	Agency/Organization	Scale	Abbreviation
2017	Comprehensive Environmental Protection Catalogue (2017 Edition)	National Development and Reform Commission (NDRC)	National	CEPC
2019	Industrial Structure Adjustment Guidance Catalogue (2019 Edition)	NDRC	National	SAGC
2020	Opinions on Further Strengthening Plastic Pollution Control	NDRC and Ministry of Ecology and Environment (MEE)	National	OFSPPC
2020	Notice on Solidly Advancing Plastic Pollution Control	NRDC and State Administration for Market Regulation	National	NSAPPC
2021	14th Five-Year Action Plan for Plastic Pollution Control	NDRC and MEE	National	FYAPPPC
2021	GB/T 40146-2021 Determination of Plastic Microbeads in Cosmetics	Standardization Administration of China under State Administration for Market Regulation	National	GB/T 40146-2021
2022	Detailed Policies for Technical Standards Development	NDRC and MEE	National	DPTSD
2023	Notice on Further Advancing Key Tasks for Plastic Pollution Control (2023–2025)	NDRC and MEE	National	NFAKTPPC
2023	Key Points of Work for Plastic Pollution Control in 2023	NDRC and MEE	National	KPWPPC

Table 2. Summary of Policies Relating to "Microbeads in Cosmetics" in the US

Year	Title	Agency / Organization	Scale	Abbreviation
1967	Fair Packaging and Labeling Act Title 15, Ch. 39.	Food & Drug Administration (FDA)	National	FPLA
2015	House Report 114-371 – Microbead-free Waters Act of 2015.	House of Representatives	National	HR 114-371

2015	Microbead-Free Waters Act.	Congress	National	MFWA
2021	Federal food, drug, and cosmetic act. Ch. 675.	Congress	National	FD&C
2022	Modernization of Cosmetics Regulation Act of 2022.	FDA	National	MoCRA

The objective is to assess the effectiveness of policies regulating the use of microplastics in PCCPs and identify areas for improvement. This analysis will deconstruct each document based on the components outlined below and evaluate the coherence of the policies in both countries up to the time of writing. The evaluation will consider the definition of key concepts, the scope of the policies, and the extent to which the instruments outlined in the guidelines have been successfully implemented. In their study of local climate adaptation, Brennan Vogel and Daniel Henstra proposed a heuristic research framework for comparative policy analysis, which identifies four fundamental elements of effective environmental policy: (i) goals, (ii) targets, (iii) instruments and (iv) agents (Vogel and Henstra). In the context of policy, (i) goals are understood as the broad normative aim or desired outcome; (ii) targets are specific aims conducive to the achievement of policy goals, commonly assigned a tangible numerical value within a measuring system; (iii) instruments are understood as the tools and mechanisms with which the policy objectives will be reached; (iv) agents are the actors involved in developing and employing the instruments for reaching these targets (Vogel and Henstra).

This study's methodological approach is exemplified through an analysis of the "Opinions on Further Strengthening Plastic Pollution Control (2020)." This document addresses all four criteria:

> **Goal**: Systematically prohibit and restrict specific plastic products, promote alternatives, standardize recycling, and establish a comprehensive management system to achieve effective plastic pollution control and support the vision of a "Beautiful China."
> **Targets**: Ban the production and sale of daily chemical products containing plastic microbeads by the end of 2020 and prohibit their sale by the end of 2022.
> **Instruments**: Develop a catalog of restricted plastic products, strengthen laws and quality standards for recycled plastics, and promote R&D on recyclable and degradable materials. Violators face legal action, and public disclosure ensures accountability.
> **Agents**: The NDRC and MEE coordinate efforts, while enterprises produce compliant alternatives, manage green systems, and improve recycling practices.

The breakdown content of documents will then be synthesized to evaluate gaps in the current framework.

Results

1. **Goals**

The United States House of Representatives introduced the Microbead-Free Waters Act (MFWA) with the purpose of banning plastic microbeads in rinse-off cosmetic products to prevent their entry into waterways across the country. Additionally, the legislation includes provisions to preempt state and local laws regarding plastic microbeads in such products. This preemption aims to address the challenges posed by a patchwork of differing state and local regulations, which complicates interstate commerce. By establishing a uniform standard, the legislation provides clarity and certainty for manufacturers and other industry stakeholders nationwide. In contrast, China has not explicitly articulated a goal specific to microplastic management in PCCPs. There is an emphasis on managing microplastics in different industries, reflecting awareness for its environmental consequence, but the keyword PCCPs only appeared a few times in a comprehensive overview compared to a specific article as in the United States (“Opinions on Further Strengthening Plastic Pollution Control” ; “14th Five-Year Action Plan for Plastic Pollution Control”; “Key Points of Work for Plastic Pollution Control in 2023).

2. **Targets**

MFWA bans “[t]he manufacture or the introduction or delivery for introduction into interstate commerce of a rinse-off cosmetic that contains intentionally added plastic microbeads” (21 USC. § 331(ddd)(1)). Note that the Federal Food Drug and Cosmetic Act, which the Act amends, defines only “cosmetic” and not “rinse-off cosmetic” (21 USC. § 321[i]). This has the potential to create ambiguity in the case of microbead-containing products that are arguably not “rinsed off.” The ban on manufacturing took effect on July 1, 2017, with the restriction on interstate commerce following a year later, on July 1, 2018 (MFWA No. 114-114, 129 Stat. 3129, § (2)(b)). The MFWA preempts state and local microbead bans unless they are identical to the federal law. China’s NDRC renewed the “Industrial Structure Adjustment Guidance Catalogue” in 2019, which classified "daily chemical products containing plastic microbeads" as third-category obsolete products. The regulations stipulated that the production of such products must cease by December 31, 2020, and their sale must be prohibited by December 31, 2022.

3. **Instruments**

Both countries regulated the manufacturing and selling of microbeads through direct source-control legislation. In relevant documents, both countries gave similar size specification and functional description of microbeads in their definition, which are specific instructions for detection and ban and thus under the category of instruments. In the US, MFWA defines plastic microbeads to mean “any solid plastic particle that is less than five millimeters in size and is intended to be used to exfoliate or cleanse the human body or any part thereof.” In China, "Detailed Standards for the Prohibition and Restriction of Related Plastic Products (2020 Edition)" of the “Notice on Steadily Promoting Plastic Pollution Control”, issued in 2020, specifies that "daily chemical products containing plastic microbeads" refer to rinse-off

cosmetics (such as shower gels, facial cleansers, scrubs, shampoos) and toothpaste or tooth powders intentionally adding solid plastic particles with a diameter of less than 5 mm to achieve effects like exfoliation, cleaning, or scrubbing. The implementation of "GB/T 40146" in 2021 clarified the definition of plastic microbeads as solid plastic particles that are smaller than or equal to 5 mm in size and insoluble.

However, a significant challenge in instruments is the lack of a unified testing standard for determining plastic microbeads in PCCPs. The primary difficulty lies in extracting plastic microbeads, particularly when organic solvents are used to dissolve certain water-insoluble matrices, which may partially dissolve the microbeads as well. Therefore, there is a need to develop a standardized and feasible detection method across the globe. Most studies rely on the Fourier-transform infrared spectroscopy (FTIR), laser diffraction, and microscopy to analyze the composition and size of microbeads. Isolating plastic microbeads often involves dissolving cosmetic products and filtering the solution to separate the microbeads. While effective for products containing a single type of microbead, this approach can struggle with more complex formulations (Kalčikova et al.). Density-based separation offers another method, where solutions of varying densities are used to isolate microbeads from other cosmetic components. Although this technique works well for products like polyethylene (PE) microbeads in facial cleansers, it may not adequately address products with mixed or more intricate compositions (Qu et al.; Habib et al.). Say for water-insoluble cosmetics, there are more difficulties with reliable detecting methods. Researchers rely on using organic solvents to disperse the product and isolate the microbeads, which can lead to potential degradation of the microbeads or interference from non-plastic compounds. Advanced tools like high-temperature gel permeation chromatography are often needed to address these issues (Hintersteiner et al.). In China, the national standard GB/T 40146-2021 employs infrared spectroscopy to qualitatively analyze certain plastic microbeads in cosmetics (excluding makeup products), but it does not support quantitative analysis. Currently, this standard is applied on a recommended, rather than mandatory, basis.

Another area of debate is whether the ban should include biodegradable microplastic. In the United States, most of the enacted state bans in cosmetic ingredients included an exemption for biodegradable plastics but did not define that term. For example, Wisconsin's law banned "synthetic plastic microbeads," defined as "any intentionally added non-biodegradable, solid plastic particle measuring less than 5 millimeters at its largest dimension that is used to exfoliate or cleanse in a product that is intended to be rinsed off" (2015 Wis. Act 43). This language appears to have been based on the Illinois statute, which contains an essentially identical definition (414 ILL. COMP. STAT. 5/52.5). California passed a different, more stringent ban that did not exempt biodegradable microbeads (CAL. PUB. RES. CODE § 42361). Instead, it defined "plastic microbead" as "an intentionally added solid plastic particle measuring five millimeters or less in every dimension" (CAL. PUB. RES. CODE § 42361) and banned the inclusion of such microbeads in personal care products, excluding prescription drugs (CAL. PUB. RES. CODE §§ 42362, 42361). Some groups argued that the "weaker" form of the state bans improperly incentivized an undesirable solution: the substitution of theoretically "biodegradable" plastics

that would not degrade under ordinary circumstances (“California Microbead Ban Closes Biodegradable Loophole”). Industry officials—and some policymakers—also objected to the state and local bans because they created a “patchwork” regulatory regime, under which the treatment of microbeads varied from state to state. While such regulatory fragmentation would not be a challenge for China—where environmental legislation is enacted at the national level—the issue of biodegradability has not been addressed in any of China’s relevant policy documents. The absence of biodegradability as a key consideration highlights a potential area for further exploration to avoid unintended consequences like those identified by critics in the United States

4. Agents

Current US laws are passed by the Congress and regulations developed by various federal agencies such as the US Environmental Protection Agency (EPA) and the US Food and Drug Administration (FDA). MFWA prohibits the manufacture, packaging, and distribution of rinse-off cosmetics containing plastic microbeads in the United States; cosmetics in the United States, except soap, do not require premarket approval from the FDA. Microplastics may be included in cosmetics without comprehensive safety assessments (Sorensen et al.). The existing legal framework in China demonstrates insufficient conceptualization and explanation of microplastics. First, the framework cannot clarify whether microplastics are considered "legal pollutants." For instance, Article 42 of the *Environmental Protection Law* enumerates the "pollutants subject to legal prevention and control", but microplastics are not included. Current lists of water pollutants in *Water Pollution Prevention and Control Law* and the *Soil Pollution Prevention and Control Law* do not include microplastics or their main components (e.g., polyethylene, polypropylene, POPs) (MEE). Consequently, microplastics cannot be classified as "hazardous waste" under the existing system, rendering risk prevention measures ineffective. Microplastics are currently regarded as a subcategory of macroplastics despite different origins. The legislation for this newly emerging pollutant relies on extension of existing policies on macroplastics. For instance, the *Soil Pollution Prevention Action Plan* is designed for concentrated and efficient soil pollution management in the context of severe pollution in agricultural and construction land. While its wording could be expanded to include microplastics, their dispersed nature and potential impact on soils conflict with the plan’s objective of efficient, concentrated governance.

Conclusion

Though microbeads from PCCPs only account for 2% of all primary microplastics released into the world's oceans according to the International Union for Conservation of Nature (IUCN) (Boucher and Friot), it is the only source that can be considered as intentional pollution, as being released as water pollutants is an inevitable part of their lifecycle and is anticipated by their manufacturers. The focus of this paper is not necessarily on the potential dangers of microbeads, but that microbeads from certain products were being directly, certainly and

unnecessarily added to the water supply (Shwartz). Regulations have been attempting to restrict the usage of microbeads in PCCPs, but a lack of specific attention for PCCPs in China, and debates about definition, scope, and detection method have led to limited effectiveness. Mindful of the fact that the breakdown of plastics already in our waterways may lead to the continued introduction of secondary microplastics, improve water filtration technologies is also a lack.

This paper explores areas for improvement and need for future research regarding goals, targets, instruments, and agents of microplastics in PCCPs regulations. With goals, both countries have an awareness for microplastic pollution. The MFWA in the US directly aimed at microbeads in PCCPs, however, microbeads in PCCP were not treated as a specific area of concern in China, sidelined as a secondary concern inside the guidelines for microplastic pollution. With targets, both countries have specific deadlines for the ban, and therefore proved to be quite effective in reducing microbeads after the bans come into effect. With instruments, both countries controlled the manufacturing and selling of microbeads through direct legislation, However, a significant challenge is the lack of a unified testing standard for determining plastic microbeads in PCCPs, and debates considering whether microplastic in PCCPs should include leave-on cosmetics and biodegradable plastic. With agents, both China and the US have multiple governmental departments working on microplastic pollution, and therefore efforts are needed to coordinate the whole lifecycle for microplastics.

As a critical content analysis employing a comparative framework to examine the policies of two countries, this study aims to identify specific areas for improvement in the United States and China. While the findings provide valuable insights for these two nations, they may not be directly generalizable to other countries. The methodology involves the manual collection and analysis of relevant policy documents, focusing on extracting key information for detailed examination. Although this approach lacks statistical outputs, such as keyword frequency analysis, it is better suited to addressing the complexity of technical terminology, which often defies simple categorization. The findings are also temporally specific, reflecting the state of microbead policies in personal care and cosmetic products (PCCPs) up to the year 2024.

The result of this paper points out futures directions for the advancement and completion of regulations regarding microbeads in PCCPs, can be used mainly by the Chinese and US government to understand current progress in the two countries. For practitioners, a key task is to acquire a better understanding of the scientific properties of microbeads and determine whether the debates around its definition, biodegradability, and existence in which types of cosmetic product categories should be regulated.

Works Cited

Akhbarizadeh, Razegheh, et al. "Reductions of Plastic Microbeads from Personal Care Products in Wastewater Effluents and Lake Waters Following Regulatory Actions." ACS ES&T Water, vol. 4, no. 2, 2024, pp. 492–499.

Andrady, Anthony L. "Microplastics in the Marine Environment." Marine Pollution Bulletin, vol. 62, no. 8, 2011, pp. 1596–1605. https://doi.org/10.1016/j.marpolbul.2011.05.030.

Arthur, C. D., Baker, J., and Bamford, H. (Eds.) Proceedings of the International Research Workshop on the Occurrence, Effects, and Fate of Microplastic Marine Debris. NOAA Technical Memorandum NOS-OR&R-30, 2008.

Bill Text—AB-888 Waste Management: Plastic Microbeads. Retrieved October 4, 2019, from https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=201520160AB88.

Boucher, J., and D. Friot. Primary Microplastics in the Oceans: A Global Evaluation of Sources. Gland: IUCN International Union for Conservation of Nature, 2017.

"California Microbead Ban Closes Biodegradable Loophole." Stormwater Report, Water Environment Federation, 8 Oct. 2015, https://stormwater.wef.org/2015/10/california-microbead-ban-closes-biodegradable-loophole/.

California Public Resources Code §§ 42361–42362. California Legislative Information, 2015, https://leginfo.legislature.ca.gov/faces/codes_displayexpandedbranch.xhtml?tocCode=PRC&division=30.

Canadian Environmental Protection Act Registry. Microbeads in Toiletries: Method 445.0. 2018. https://www.canada.ca/en/environment-climate-change/services/canadian-environment-protection-act-registry/publications/microbeads-toiletries-method-445-0.html.

Chen, Gui-rong, Li Neng, and Xu Yu-xun. "Regulatory Status of Plastic Microbeads Used in Cosmetic Products at Home and Abroad." *Intertek Testing Services Shenzhen Co., Ltd.*, vol. 41, no. 12, Dec. 2018, pp. 14–18. DOI:10.13222/j.cnki.dc.2018.12.005.

Chen, Guofu and Qihang Li. "The Cognitive Foundation of Law, Endogenous Preferences, and the Choice of Rules for Rights Protection" [J]. *Nankai Journal (Philosophy and Social Sciences Edition)*, 2013(4): 139–147. n.d.

Cheung, Pui Kwan, and Lincoln Fok. "Characterisation of Plastic Microbeads in Facial Scrubs and Their Estimated Emissions in Mainland China." Water Research, vol. 122, 2017, pp. 53–61. https://doi.org/10.1016/j.watres.2017.05.053.

"China Skincare Market Size, Share & COVID-19 Impact Analysis." Fortune Business Insights, www.fortunebusinessinsights.com/china-skincare-market-107629. Accessed 10 Dec. 2024.

Cole, M., Lindeque, P., Halsband, C., and Galloway, T. S. "Microplastics as Contaminants in the Marine Environment: A Review." Marine Pollution Bulletin, vol. 62, 2011, pp. 2588–2597.

Congress. Federal Food, Drug, and Cosmetic Act. Ch. 675 of the 75th Congress, Section 302, 2021. https://www.govinfo.gov/content/pkg/COMPS-973/pdf/COMPS-973.pdf.

Doughty, Rachel, and Marcus Eriksen. "The Case for a Ban on Microplastics in Personal Care Products." Tulane Environmental Law Journal, vol. 27, 2014, pp. 277–278.

E, United Nations Environment Programme. Plastic in cosmetics: Are we polluting the environment through our personal care? plastic ingredients that contribute to marine microplastic litter[R]. The Global Programme of Action for the Protection of the Marine.

FDA. The Microbead-Free Waters Act: FAQs. 2020. Retrieved February 1, 2022, from https://www.fda.gov/cosmetics/cosmetics-laws-regulations/microbead-free-waters-act-faqs.

Fiorino, Daniel J. Making Environmental Policy. Berkeley: University of California Press, 1995.

Galafassi, Silvia, Luca Nizzetto, and Pietro Volta. "Plastic Sources: A Survey across Scientific and Grey Literature for Their Inventory and Relative Contribution to Microplastics Pollution in Natural Environments, with an Emphasis on Surface Water." Science of the Total Environment, vol. 693, 2019. https://doi.org/10.1016/j.scitotenv.2019.07.305.

Gawlik, B.M., Głowacka, N., Feldman, D.L. et al. The scientist, the politician, the artist and the citizen: how water united them. Environ Sci Eur 30, 12 (2018). https://doi.org/10.1186/s12302-018-0141-5

Graney, Guy. "Slipping Through the Cracks: How Tiny Plastic Microbeads Are Currently Escaping Water Treatment Plants and International Pollution Regulation." Fordham International Law Journal, vol. 39, 2016, pp. 1023.

Gregory, M. R. "Plastic 'Scrubbers' in Hand Cleansers: A Further (and Minor) Source for Marine Pollution Identified." Marine Pollution Bulletin, vol. 32, 1996, pp. 867–871.

Habib, R. Z., Salim Abdoon, M. M., Al Meqbaali, R. M., et al. "Analysis of Microbeads in Cosmetic Products in the United Arab Emirates." Environmental Pollution, vol. 258, 2020, 113831.

Hintersteiner, I., Himmelsbach, M., and Buchberger, W. W. "Characterization and Quantitation of Polyolefin Microplastics in Personal-Care Products Using High-Temperature Gel-Permeation Chromatography." Analytical & Bioanalytical Chemistry, vol. 407, no. 4, 2015, pp. 1253–1259.

Huang, Kaisheng, Chen Zehong, Liao Jia, et al. "Study on the Identification Methods and Risk Investigation of Plastic Microbeads in Daily Chemical Products." Guangdong Chemical Industry, vol. 48, no. 10, 2021, pp. 69–71, 68.

Illinois General Assembly—Full Text of Public Act. No. Retrieved December 10, 2024, from http://www.ilga.gov/legislation/publicacts/fulltext.asp?Name=098-0638.

Kalčíková, G., Alič, B., Skalar, T., et al. "Wastewater Treatment Plant Effluents as Source of Cosmetic Polyethylene Microbeads to Freshwater." *Chemosphere*, vol. 188, 2017, pp. 25-31.

Karami, Ali, Golieskardi, Adeleh, Choo, Chee Kong, et al. "A High-Performance Protocol for Extraction of Microplastics in Fish." *Science of the Total Environment*, vol. 578, 2017, pp. 485-494.

Kim, Seung-Kyu, and Seung-Hyun Lim. "Microbeads in Cosmetics as Emerging Pollutants of Concern." *Marine Pollution Bulletin*, vol. 104, no. 1–2, 2016, pp. 302-310.

Kutralam-Muniasamy, Gunasekaran, Perez-Guevara, Fabiola, Elizalde-Martinez, Ireri, et al. "Quantitative Assessment of Microplastics and Their Characteristics in the Consumer Facial Cleansers Marketed in Mexico." *Environmental Pollution*, vol. 236, 2018, pp. 599-608.

Lechner, Andreas, and Thomas Ramler. "The Discharge of Certain Amounts of Industrial Microbeads to a Municipal Wastewater Treatment Plant." *Science of the Total Environment*, vol. 607–608, 2017, pp. 761-766.

Leslie, Heather A., et al. "Microplastics En Route: Field Measurements in the Dutch River Delta and Amsterdam Canals, Wastewater Treatment Plants, North Sea Sediments, and Biota." *Environmental International*, vol. 101, 2017, pp. 133-142.

Li, J., Liu, H., and Paul Chen, J. "Microplastics in Freshwater Systems: A Review on Occurrence, Environmental Effects, and Methods for Microplastics Detection." *Water Research*, vol. 137, 2018, pp. 362-374.

Magni, Stefano, et al. "An Overview of the Environmental Applicability of Microplastics and Nanoplastics." *Ecotoxicology*, vol. 29, no. 2, 2020, pp. 34-50.

Mason, Sherri A., Vanden Berg, Jake, and Sutton, Rebecca. "Occurrence of Microplastics in the United States: A Comparison of Wastewater, Source Waters, and Treated Drinking Water." *Environmental Pollution*, vol. 246, 2019, pp. 608-618.

Ministry of Environmental Protection. Comprehensive Environmental Protection Catalogue (2017 Edition) (环境保护综合名录（2017 年版)) [EB/OL]. No. (2018-02-06). [Accessed: February 6, 2018]. Available at: https://www.mee.gov.cn/gkml/sthjbgw/qt/201802/t20180206_430933.htm.

Mintenig, S. M., Löder, M. G. J., Primpke, S., and Gerdts, G. "Microplastics in Drinking Water: Field Testing and Analytical Challenges." *Current Opinion in Environmental Science & Health*, vol. 7, 2019, pp. 15-21.

Napper, Imogen E., and Richard C. Thompson. "Release of Synthetic Microplastic Plastic Fibres from Domestic Washing Machines: Effects of Fabric Type and Washing Conditions." *Marine Pollution Bulletin*, vol. 112, no. 1–2, 2016, pp. 39-45.

National Development and Reform Commission (NDRC). Comprehensive Environmental Protection Catalogue (2017 Edition). 2017.

National Development and Reform Commission (NDRC). Industrial Structure Adjustment Guidance Catalogue (2019 Edition). 2019.

National Development and Reform Commission (NDRC), and Ministry of Ecology and Environment (MEE). Opinions on Further Strengthening Plastic Pollution Control. 2020.

National Development and Reform Commission (NDRC), and State Administration for Market Regulation. Notice on Solidly Advancing Plastic Pollution Control. 2020.

National Development and Reform Commission (NDRC), and Ministry of Ecology and Environment (MEE). 14th Five-Year Action Plan for Plastic Pollution Control. 2021.
Standardization Administration of China under State Administration for Market Regulation. GB/T 40146-2021 Determination of Plastic Microbeads in Cosmetics. 2021.
National Development and Reform Commission (NDRC), and Ministry of Ecology and Environment (MEE). Detailed Policies for Technical Standards Development. 2022.
National Development and Reform Commission (NDRC), and Ministry of Ecology and Environment (MEE). Notice on Further Advancing Key Tasks for Plastic Pollution Control (2023–2025). 2023.
National Development and Reform Commission (NDRC), and Ministry of Ecology and Environment (MEE). Key Points of Work for Plastic Pollution Control in 2023. 2023.
Standing Committee of the National People's Congress. Product Quality Law of the People's Republic of China (中华人民共和国产品质量法) [EB/OL]. . No. (2019-01-07). [Accessed: December 10, 2024]. Available at: http://www.npc.gov.cn/npc/c2/c30834/201905/t20190521_296662.html. . n.d.
National Oceanic and Atmospheric Administration (NOAA). *Microplastics: Occurrence, Sources, and Effects.* NOAA Technical Memorandum NOS-OR&R-37, 2015.
OECD. "Improving Plastics Management: Trends, Policy Responses, and the Role of International Cooperation and Trade." *OECD Environment Policy Papers*, no. 12, 2018. https://doi.org/10.1787/c5f7c448-en.
Peng, Linlin, et al. "Microplastics Contamination in Freshwater Systems: A Review of Occurrence, Interactions, and Management Strategies." *Science of the Total Environment*, vol. 586, 2017, pp. 89-96.
Plastic Soup Foundation. "Plastic Microbeads in Cosmetics." Retrieved January 12, 2025, from https://www.plasticsoupfoundation.org/en/plastic-microbeads-in-cosmetics/.
Rochman, Chelsea M., et al. "Policy: Classify Plastic Waste as Hazardous." *Nature*, vol. 494, 2013, p. 169.
Ryan, Peter G., et al. "Trends in Coastal Plastic Debris: Monitoring the 1980s versus 1990s at the Same Beaches." *Marine Pollution Bulletin*, vol. 40, 2000, pp. 121–126.
Steensgaard, Ida M., et al. "From Macro- to Microplastics: Analysis of EU Regulation along the Life Cycle of Plastic Bags." *Environmental Science & Policy*, vol. 69, 2017, pp. 10–20.
Suaria, Giuseppe, Stefano Aliani, Silvia Merlino, and Marinella Abbate. "The Occurrence of Paraffin and Other Petroleum Waxes in the Marine Environment: A Review of the Current Legislative Framework and Shipping Operational Practices." Frontiers in Marine Science, vol. 5, 2018, doi:10.3389/fmars.2018.00094. Accessed at https://www.frontiersin.org/journals/marine-science/articles/10.3389/fmars.2018.00094.
Sun, Qing, Shu-Yan Ren and Ni Hong-Gang. ""Incidence of Microplastics in Personal Care Products: An Appreciable Part of Plastic Pollution." Science of the Total Environment, vol. 742, 2020, 140218. ScienceDirect, doi:10.1016/j.scitotenv.2020.140218. ." (n.d.).

Sutton, Rebecca, et al. "Reducing Microplastics in San Francisco Bay through Policies Targeting Wastewater Treatment Plants and Textile Waste." *Environmental Science & Technology Letters*, vol. 6, no. 2, 2019, pp. 85–91.

Thompson, Richard C., et al. "Lost at Sea: Where Is All the Plastic?" *Science*, vol. 304, no. 5672, 2004, pp. 838.

Van Cauwenberghe, Lisbeth, and Colin R. Janssen. "Microplastics in Bivalves Cultured for Human Consumption." *Environmental Pollution*, vol. 193, 2014, pp. 65-70.

Wagner, Martin, and Scott Lambert. "Freshwater Microplastics: Emerging Environmental Contaminants?" *Springer International Publishing*, 2018.

World Health Organization. "Dietary and inhalation exposure to nano- and microplastic particles and potential implications for human health. Geneva: World Health Organization; 2022. Licence: CC BY-NC-SA 3.0 IGO."

Wright, Stephanie L., and Frank J. Kelly. "Plastic and Human Health: A Micro Issue?" *Environmental Science & Technology*, vol. 51, 2017, pp. 6634–6647.

Zhang, Chenchen, et al. "Occurrence and Removal Efficiencies of Microplastics in Wastewater Treatment Plants in Coastal Cities of China." *Environmental Pollution*, vol. 248, 2019, pp. 115–122.

Zhang, Wei, et al. "Brief Reviews of the Control of Plastic Microbeads in Cosmetics." 12 April 2024.

Burnout and Happiness in High School Student Athletes: A Systematic Review
By Anatoli Monsalve

Abstract

This study examines the relationship between sports-related burnout and happiness in high school student-athletes who navigate the dual demands of academics and athletics during critical identity development. Existing research primarily focuses on collegiate and professional athletes, leaving a gap in understanding how burnout manifests in younger populations. This study synthesizes findings from academic databases through a systematic literature review, analyzing different risk factors such as perfectionism, intense demands, adverse coaching climates, and early specialization. Results highlight that burnout arises from a mismatch between demands and resources, leading to detrimental short-term and long-term consequences, including reduced performance, mental health challenges, and heightened dropout rates. Addressing these risks requires creating supportive environments, balancing demands, and emphasizing intrinsic motivation.

Introduction

Burnout is a state of physical, emotional, and mental exhaustion caused by prolonged stress (Maslach). This could impact perceived happiness levels and a sense of well-being and contentment (Prentice et al.). This is especially concerning regarding the high school student-athlete population, which encounters competing demands navigating the demands of academics and athletics during a critical development period, like identity formation, which poses unique challenges (Lee et al.). Currently, extant literature focuses on sports-related burnout and happiness, focusing on older samples, such as collegiate or professional athletes, because they notoriously face immense pressure in and out of the respective sport. However, this leaves a gap in our understanding of high school student-athletes: how do these issues manifest in high school populations? Teenage years are a crucial time for identity development, and being an athlete and all the pressures and glories can also play a significant role in shaping it.

Furthermore, the literature defines sports-related burnout in high school athletes differently than in collegiate and professional populations. Burnout in collegiate and professional populations was more defined by the pressure of widely broadcast competitions, whereas that is not a typical factor for high school populations.

Methods

To address the research question, I am conducting a systematic literature review. Specifically, this method involves collecting, analyzing, and synthesizing data from relevant studies. First, I will search academic databases such as Google Scholar. I will use relevant keywords (e.g., “burnout,” “athlete burnout,” “happiness,” and “high school athletes”). Next, I will read the titles and abstracts to exclude irrelevant studies. Irrelevant studies are those that do not pertain to the relationship between sports-related-burnout, its risk factor, and its

consequences in high-school athletes. Then, reading the whole paper will ensure that the remaining studies are relevant and lack bias. From there, I will extract data where I will capture the study characteristics (e.g., authors, year, location), participant details (e.g., sample size, age, gender), methods (e.g., study design, measures of burnout and happiness), and key findings (e.g., relationships between burnout and happiness). Then, I will analyze and synthesize this information to identify patterns in how sports-related burnout affects happiness. Finally, I will identify gaps in the current discourse and propose another area of study.

1. Risk Factors

High Expectations and Perfectionism

While high expectations serve as motivation at times, when externally imposed expectations are paired with maladaptive perfectionism, the likelihood of burnout increases. Sorkkila et al. (2017) employed structural equation modeling on 391 student-athletes and 448 parents and found that external success expectations—especially from parents—correlated with higher burnout risk in high school athletes (Sorkkila, Aunola, et al.). The pressure to meet expectations from those close to us often leads to the fear of failure, resulting in athletes doubting their skills and ultimately feeling demotivated if they do not reach their goals. Maladaptive perfectionism further increases this risk of burnout. Chen et al. (2009) found that athletes who exhibited "adaptive perfectionism" (a drive to excel without being crippled by mistakes) were less likely to experience burnout (Chen et al.). Conversely, athletes exhibiting "maladaptive perfectionism" (constant self-criticism and the inability to accept setbacks) were more prone to burnout. Therefore, the external pressures, in conjunction with maladaptive perfectionism, create a toxic environment where athletes may associate how they perform based upon the conditional approval from parents and coaches. In turn, it substantially heightens the risk of burnout by undermining athletes' intrinsic enjoyment of the sport.

Intense Demands

The demands of intense training and competition contribute to burnout among high-school athletes, particularly among juniors and seniors (Chu et al.). This increased risk during the "investment years" (ages 16 and above) of the Developmental Model of Sport Participation stems from higher training loads, more significant competition pressure, early sport specialization, and limited recovery time. Furthermore, high school athletes also face additional academic pressures while navigating through the increasing demands of the sport (Appleton et al.)(H. Gustafsson et al.). All these factors overwhelm many athletes, leading to sports-related burnout as they have to balance more rigorous training and academic schedules in these "investment years"(H. Gustafsson et al.).

To further explain this dilemma, Goodger et al. (2007) note that the pursuit of performance—both athletically and academically—can trap athletes. Athletes are left in a cycle that leaves little room for recovery or other aspects of a balanced life, ultimately increasing their

risk of burnout (Goodger et al., “Burnout in Sport”). Gustafsson et al. (2007) expand on this idea by arguing that the increasing professionalization of youth sports, as noted by Gould and Whitley (2009) as well, places excessive pressure on high school athletes already navigating the demands of academics and social development(Gould and Whitley)(Henrik Gustafsson, Kenttä, et al.). Thus, the intensity of training and competition they go through increases high school athletes’ vulnerability to burnout.

Negative Coaching Climate

A negative coaching climate significantly increases the risk of sports-related burnout among high school athletes. Into et al. (2020) highlight how demotivating coaching climates—those that have controlling behaviors, lack of autonomy support, and ineffective communication—create psychologically damaging environment (Into et al.). Coaches who fail to build trust in the training or do not allow athletes to voice their concerns contribute to an adverse climate(Into et al.)(Dannis). This climate typically fosters emotional exhaustion: no matter what the athlete does, they do not feel any control or trust in their training; instead, they are just going through the motions(Granz et al.)(Chu and Zhang). As such, athletes in these climates often feel controlled and undervalued. This sense of demotivation is closely tied to feelings of inadequacy and overwhelming pressure(Westfall et al.). Therefore, a positive coaching climate that fosters trust, autonomy, and effective communication is crucial in preventing burnout among high school athletes.

Another way in which coaches can increase the risk of sports-related burnout is when they focus solely on winning or immediate results. In this vein, negative coaching involves the coach focusing solely on criticizing the athlete without positively reinforcing them when they do something correctly(Henrik Gustafsson, Hassmén, et al.). As a result, this creates an environment fueled by fear because, typically, athletes seek the coaches' approval(Mageau and Vallerand). In this environment, athletes often fear failure and feel pressured to meet their coach’s unrealistic expectations, decreasing their intrinsic motivation and enjoyment of the sport—it is not about themselves but about pleasing their coaches(Westfall et al.). Consequently, this increases their vulnerability to burnout.

Demands-Resources Imbalance

Burnout in high school athletes often stems from an imbalance as the ‘demands’ placed on them exceed the ‘resources’ available to manage these demands(Sorkkila, Aunola, et al.; Chu et al.; Into et al.; Granz et al.; Sorkkila, Ryba, et al.). When demands (such as academic expectations, intensive training schedules, or pressure of competition) exceed an athlete’s resources, the risk of emotional, physical, and mental exhaustion increases significantly which then heighten the likelihood of feeling burnout(Sorkkila, Aunola, et al.; Chu et al.; Into et al.; Granz et al.; Sorkkila, Ryba, et al.). More specifically, Gustafsson et al. (2007) emphasize that when there are numerous pressures put on the athlete (for example, high-level athletic

performance and academic responsibilities) it can overwhelm young athletes, especially when they lack adequate coping mechanisms or support systems(Henrik Gustafsson, Kenttä, et al.).

The availability of resources (strong social support networks, effective coping strategies, and a sense of autonomy in their athletic careers) is crucial in buffering athletes against burnout. Chu et al. (2022) found that athletes with access to a supportive network of family, friends, coaches, and teammates are better equipped to handle the pressures of their roles as students and athletes(Chu et al.). Additionally, Goodger et al. (2007) highlights that developing effective coping mechanisms (stress management techniques and positive self-talk) can lessen the pressures athletes experience while managing their academic and athletic lives(Goodger et al., "Burnout in Sport"). Furthermore, Markati et al. (2018) emphasize the importance of maintaining an "optimum training zone," suggesting that athletes should achieve peak performance without overexertion as this reduces the risk of burnout(Markati et al.).

Early Specialization

Early specialization in sports increases the risk of burnout among high school athletes. While there are benefits to focused training, an emphasis on specialization, particularly during the critical developmental stages of adolescence, can create an imbalance between demands and resources, negatively impacting athletes' well-being and potentially leading to burnout(Brenner). Chu et al. (2022) explain this progression using the Developmental Model of Sport Participation, which outlines how athletes move from the "sampling years" of exploring multiple sports to the "specializing years" and, ultimately, the "investment years" of intense focus on a single sport(Chu et al.). However, prematurely pushing into the "investment years" can lead to an increased risk of burnout. When young athletes exclusively focus on a single sport, they have fewer opportunities to engage in other activities, develop diverse interests, and build relationships outside of their sport, leading to a sense of identity that is overly dependent on athletic achievement(Coakley). This can lead to social isolation and a sense of being "trapped" by their sport as they overly rely on athletic identity(Markati et al.). So, when challenges or setbacks arise, it makes younger athletes more vulnerable to burnout. After all, if one's identity is heavily tied to one's sport and one does not view oneself as "good enough," it makes one more susceptible to being burnt out(Appleton et al.; Henrik Gustafsson, Hassmén, et al.).

Furthermore, early specialization (which emphasizes repetitive training and year-round competition) can increase the risk of overuse injuries, especially in young athletes whose bodies are still developing (Brenner). Markati et al. (2019) highlight that the inability to perform due to injury can impact an athlete's self-worth and lead to feelings of inadequacy(Markati et al.). This is incredibly challenging as most youth athletes who specialize early tie their identity heavily to their sport (Appleton et al.; Henrik Gustafsson, Hassmén, et al.; Coakley). These feelings of inadequacy increase the risk of burnout (Sorkkila, Ryba, et al.). Additionally, athletes may experience pressure (whether real or perceived) to return to their—increasing the risk of another injury while adding more stress to their existing amount and ultimately manifesting as feeling burnt out (Wilczyńska et al.).

2. Consequences

Short-Term Consequences

One of the short-term consequences of sports-related burnout is a decline in athletic performance (Lee et al.; Henrik Gustafsson, Hassmén, et al.). When athletes experience burnout, they suffer from physical and emotional exhaustion, which depletes their energy and reduces their ability to perform at their peak (Lee et al.). Exhaustion manifests as physical fatigue and emotional detachment, leading to diminished effort during training and competition(Chu et al.). Sometimes, due to burnout athletes are unable to maintain the same intensity or endurance they once could achieve which contributes to worse performances (Dannis). Athletes struggling with burnout no longer feel the same enthusiasm or passion for their sport, significantly diminishing their intrinsic drive to improve and compete(Moen et al.). The enjoyment athletes once felt from their sport that fueled their commitment to keep pushing through difficult times becomes overshadowed by feelings of fatigue and indifference stemming from burnout; thus, this makes it difficult for athletes to push themselves to their maximum capacity during practice or games (Moen et al.).

Burnout affects athletes' mental abilities, such as focus, critical components of successful athletic performance(Lee et al.). As burnout intensifies, athletes experience mental fatigue, which impairs an athlete’s ability to concentrate and respond to unexpected situations which are common in competitions(Chu et al.). This loss of focus can lead to inconsistent performance because when athletes struggle to maintain attention to specific things, they start to make errors they would typically avoid(Chu et al.). The effect of exhaustion and lack of focus makes it difficult for athletes to achieve the consistency necessary for success in their sport, further deteriorating their confidence and fueling the burnout cycle(Henrik Gustafsson, Kenttä, et al.; Wilczyńska et al.).

Burnout diminishes the enjoyment athletes once felt from their sport. Burnout leads athletes to view their sport negatively, associating it more with anxiety and pressure than with the fulfillment they once experienced(Coakley). With this internal shift from intrinsic to extrinsic motivation—where athletes feel driven by external pressures from parents or coaches rather than their own personal satisfaction—athletes lose their sense of autonomy in relation to their sport(Dannis).

As burnout develops, it also influences athletes’ social behaviors and relationships. Athletes may begin to withdraw from their teammates, coaches, and other social interactions within the sports environment. Feelings of cynicism and detachment toward their sport make these interactions seem unenjoyable, leading athletes to avoid team activities or social events(Dannis). Moreover, as athletes’ performance declines, they may fear judgment or criticism from those close to them (peers, parents, and coaches); this can drive them to retreat from social interactions to avoid being unfavorably compared to their peers (Dannis).

Sports-related burnout can also spill over into academics, causing high school athletes to struggle with their schoolwork(Sorkkila, Ryba, et al.). The mental and emotional exhaustion that

stems from burnout impairs an athlete's ability to focus in class, retain information, and complete assignments (Into et al.). Emotional detachment from their sport can translate into a similar disinterest in academics, reducing their motivation to study or engage with schoolwork (Dannis). Burnout's impact on mental energy leaves athletes with little energy to handle all of the pressures they face, from school or from their sport (Cosh and Tully). Moreover, the time constraints imposed by intense training schedules limit the time athletes can dedicate to their studies, this further exacerbates the challenge of balancing both athletic and academic responsibilities (Into et al.; Cosh and Tully). Therefore, combined with the fact that burnout leaves athletes feeling demotivated, their limited time to complete these assignments makes them prioritize sports over education. Student-athletes may talk about prioritizing sport over academics as a way to overcome stressors (Cosh and Tully). In addition, student-athletes sometimes say that they only need to pass their classes because they perceive the sports domain as more prominent than school (Ryba et al.).

Long-Term Consequences

The long-term consequences of sports-related burnout in high school athletes impact their future athletic involvement, mental health, academic and social development, and overall well-being.

A long-term consequence is the increased likelihood of sports dropout (Lee et al.). Athletes experiencing burnout often develop negative perceptions of their involvement in sports as they continue to doubt their abilities in the wake of prolonged stress(Dannis). The prolonged stress and dissatisfaction accompanying burnout can solidify a view of sports as a source of distress rather than enjoyment (Moen et al.). This sustained negative effect makes continued participation less appealing, with athletes gradually disengaging physically and emotionally (Into et al.). Burnout also erodes an athlete's "athletic identity" - a central component of many high schoolers' sense of self. As athletes lose their sense of accomplishment and competence in their sport, they may feel less connected to the identity they once derived from their athletic involvement. In some cases, this could lead to seeking alternative areas of success and belonging where they feel more "valued" (academics, social clubs, or hobbies) as they seek fulfillment outside the pressures of competitive sports.

Burnout's emotional toll can also result in long-lasting negative feelings about sports participation, which may extend well into adulthood. The intense pressures and sense of failure associated with burnout can lead to emotional trauma if left unaddressed (Gould and Whitley). Athletes who experience burnout often carry unresolved feelings of inadequacy and stress (Dannis). Looking long-term, this can influence how they view their past involvement in sports and shape their willingness to engage in future physical activities. In addition, as burnout often significantly impacts self-esteem since athletes associate their perceived lack of achievement with personal failure this may foster insecurities in sports and other areas of life (Sorkkila, Ryba, et al.). These negative self-perceptions can be long-lasting, affecting the individual's confidence in multiple different enviornments. Moreover, the painful experiences associated with burnout

may lead to avoidance of similar situations in the future, with former athletes potentially steering clear of competitive sports or physical activities where they fear experiencing the same stress and failure(Henrik Gustafsson, Kenttä, et al.).

Burnout also impacts mental health in the long run. Burnout can exacerbate or even trigger mental health challenges, mainly if the emotional stress and lack of support go unaddressed (Lee et al.). Chronic stress (prolonged exposure to stress hormones such as cortisol) disrupts brain chemistry, increasing athletes' vulnerability to conditions like anxiety and depression. Burnout often results from unmet psychological needs for autonomy, competence, and relatedness—critical factors for psychological well-being(Appleton et al.). When these needs are not fulfilled, athletes may experience a sense of hopelessness, leading to long-term mental health struggles(Moen et al.). In addition, if athletes do disengage from sports, it means they may lose buffers to mental health problems such as social support, regular physical activity, and a sense of purpose(Lee et al.). These 'buffers' can help mitigate the risk of mental health issues as without them athletes are more vulnerable to emotional difficulties.

Burnout can also impede athletes' academic and career development (Sorkkila, Ryba, et al.; Lee et al.). Athletes experiencing burnout often face academic challenges due to the cognitive and emotional demands of burnout, which spill over into the classroom (Sorkkila, Ryba, et al.). The lowered academic performance often accompanying burnout can have cascading effects, limiting opportunities for higher education and potentially impairing future career prospects (Ryba et al.). Burnout can cause athletes to fall behind in their coursework, so thinking long-term can create gaps in learning that are difficult to recover from. The intense focus on coping with athletic and academic stress may also limit athletes' engagement in career exploration, reducing their chances of gaining valuable experience or developing interests outside of sports(Sorkkila, Ryba, et al.).

Finally, burnout has potential long-term health consequences. Chronic stress disrupts the body's ability to regulate stress hormones, leading to an altered stress response system that increases the risk of developing chronic conditions such as cardiovascular disease or mental health disorders(Coakley; Lee et al.). Athletes may also adopt unhealthy coping mechanisms to manage their stress, such as substance use or disordered eating, which can have serious long-term effects on their physical and emotional well-being (Henrik Gustafsson, Hassmén, et al.). Burnout can disrupt sleep patterns and lead to chronic fatigue, reducing energy levels and encouraging sedentary behaviors, undermining health and fitness (Moen et al.; Lee et al.). Although more research is needed to fully understand the long-term health impacts of burnout, the chronic stress experienced by athletes suggests a potential for lasting adverse outcomes.

Conclusion

The current literature on sports-related burnout in high school athletes reveals numerous short-term and long-term consequences, highlighting the urgent need for targeted interventions. Burnout leads to reduced athletic performance, increased injury risk, and academic struggles in the short term. At the same time, long-term effects include sports dropout, mental health challenges, and difficulties in personal development. Therefore, as burnout can manifest itself

early on, it is essential to identify it early and prevent it, particularly in the high school population, when burnout can impact athletic performance and broader areas of life.

A limitation of the current literature is that most studies are cross-sectional; therefore, as cross-sectional studies provide a snapshot of how burnout evolves, there is a lack of understanding of how burnout affects high-school athletes over time. One of the most pressing implications is that high school athletes must move beyond focusing on performance metrics alone. High school athletes often experience intense physical and emotional demands, and burnout tends to go unnoticed until their performance declines. An approach that regularly assesses high school athletes' emotional well-being and psychological stress is essential to address this. Interventions should focus on reducing burnout symptoms and minimizing risk factors, reducing the risk of burnout before it becomes severe.

Works Cited

Appleton, Paul R., et al. "Relations between Multidimensional Perfectionism and Burnout in Junior-Elite Male Athletes." *Psychology of Sport and Exercise*, vol. 10, no. 4, July 2009, pp. 457–65. *DOI.org*, https://doi.org/10.1016/j.psychsport.2008.12.006.

Brenner, Joel S. *Overuse Injuries, Overtraining, and Burnout in Child and Adolescent Athletes*.

Chen, Lung Hung, et al. "An Examination of the Dual Model of Perfectionism and Adolescent Athlete Burnout: A Short-Term Longitudinal Research." *Social Indicators Research*, vol. 91, no. 2, Apr. 2009, pp. 189–201. *DOI.org*, https://doi.org/10.1007/s11205-008-9277-9.

Chu, Tsz Lun (Alan), et al. "Developmental Differences in Burnout Among High School Athletes in the United States: A Gendered Perspective." *Journal of Clinical Sport Psychology*, vol. 16, no. 1, Mar. 2022, pp. 42–54. *DOI.org*, https://doi.org/10.1123/jcsp.2021-0017.

Chu, Tsz Lun (Alan), and Tao Zhang. "The Roles of Coaches, Peers, and Parents in Athletes' Basic Psychological Needs: A Mixed-Studies Review." *International Journal of Sports Science & Coaching*, vol. 14, no. 4, Aug. 2019, pp. 569–88. *DOI.org* , https://doi.org/10.1177/1747954119858458.

Coakley, Jay. "Burnout among Adolescent Athletes: A Personal Failure or Social Problem?" *Sociology of Sport Journal*, vol. 9, no. 3, Sept. 1992, pp. 271–85. *DOI.org* , https://doi.org/10.1123/ssj.9.3.271.

Cosh, Suzanne, and Phillip J. Tully. "Stressors, Coping, and Support Mechanisms for Student Athletes Combining Elite Sport and Tertiary Education: Implications for Practice." *The Sport Psychologist*, vol. 29, no. 2, June 2015, pp. 120–33. *DOI.org* , https://doi.org/10.1123/tsp.2014-0102.

Dannis, Taylor. "Athletic Burnout in High School Students: A Social Perspective." *Journal of Student Research*, vol. 12, no. 3, Aug. 2023. *DOI.org* , https://doi.org/10.47611/jsrhs.v12i3.4891.

Goodger, Kate, et al. "Burnout in Sport: A Systematic Review." *The Sport Psychologist*, vol. 21, no. 2, June 2007, pp. 127–51. *DOI.org* , https://doi.org/10.1123/tsp.21.2.127.

---. "Burnout in Sport: A Systematic Review." *The Sport Psychologist*, vol. 21, no. 2, June 2007, pp. 127–51. *DOI.org* , https://doi.org/10.1123/tsp.21.2.127.

Gould, Daniel, and Meredith A. Whitley. "Sources and Consequences of Athletic Burnout among College Athletes." *Journal of Intercollegiate Sport*, vol. 2, no. 1, June 2009, pp. 16–30. *DOI.org* , https://doi.org/10.1123/jis.2.1.16.

Granz, Hanna L., et al. "Risk Profiles for Athlete Burnout in Adolescent Elite Athletes: A Classification Analysis." *Psychology of Sport and Exercise*, vol. 41, Mar. 2019, pp. 130–41. *DOI.org* , https://doi.org/10.1016/j.psychsport.2018.11.005.

Gustafsson, H., et al. "Fear of Failure, Psychological Stress, and Burnout among Adolescent Athletes Competing in High Level Sport." *Scandinavian Journal of Medicine & Science in Sports*, vol. 27, no. 12, Dec. 2017, pp. 2091–102. *DOI.org* , https://doi.org/10.1111/sms.12797.

Gustafsson, Henrik, Peter Hassmén, et al. "A Qualitative Analysis of Burnout in Elite Swedish Athletes." *Psychology of Sport and Exercise*, vol. 9, no. 6, Nov. 2008, pp. 800–16. *DOI.org* , https://doi.org/10.1016/j.psychsport.2007.11.004.

Gustafsson, Henrik, Göran Kenttä, et al. “Prevalence of Burnout in Competitive Adolescent Athletes.” *The Sport Psychologist*, vol. 21, no. 1, Mar. 2007, pp. 21–37. *DOI.org* , https://doi.org/10.1123/tsp.21.1.21.

Into, Sonja, et al. “Relationship between Coaching Climates and Student-Athletes’ Symptoms of Burnout in School and Sports.” *Sport, Exercise, and Performance Psychology*, vol. 9, no. 3, Aug. 2020, pp. 341–56. *DOI.org* , https://doi.org/10.1037/spy0000180.

Lee, Keunchul, et al. “Relationships Among Stress, Burnout, Athletic Identity, and Athlete Satisfaction in Students at Korea’s Physical Education High Schools: Validating Differences Between Pathways According to Ego Resilience.” *Psychological Reports*, vol. 120, no. 4, Aug. 2017, pp. 585–608. *DOI.org* , https://doi.org/10.1177/0033294117698465.

Mageau, Geneviève A., and Robert J. Vallerand. “The Coach–Athlete Relationship: A Motivational Model.” *Journal of Sports Sciences*, vol. 21, no. 11, Nov. 2003, pp. 883–904. *DOI.org* , https://doi.org/10.1080/0264041031000140374.

Markati, Alexandra, et al. “Psychological and Situational Determinants of Burnout in Adolescent Athletes.” *International Journal of Sport and Exercise Psychology*, vol. 17, no. 5, Sept. 2019, pp. 521–36. *DOI.org* , https://doi.org/10.1080/1612197X.2017.1421680. *Maslach*.

Moen, Frode, et al. *Examining Possible Relationships between Mindfulness, Stress, School- and Sport Performances and Athlete Burnout*.

Prentice, Shaun, et al. “Burnout, Wellbeing and How They Relate: A Qualitative Study in General Practice Trainees.” *Medical Education*, vol. 57, no. 3, Mar. 2023, pp. 243–55. *DOI.org*, https://doi.org/10.1111/medu.14931.

Ryba, Tatiana V., et al. “A New Perspective on Adolescent Athletes’ Transition into Upper Secondary School: A Longitudinal Mixed Methods Study Protocol.” *Cogent Psychology*, edited by Nikos Zourbanos, vol. 3, no. 1, Dec. 2016, p. 1142412. *DOI.org*, https://doi.org/10.1080/23311908.2016.1142412.

Sorkkila, Matilda, Kaisa Aunola, et al. “A Person-Oriented Approach to Sport and School Burnout in Adolescent Student-Athletes: The Role of Individual and Parental Expectations.” *Psychology of Sport and Exercise*, vol. 28, Jan. 2017, pp. 58–67. *DOI.org*, https://doi.org/10.1016/j.psychsport.2016.10.004.

Sorkkila, Matilda, Tatiana V. Ryba, et al. “Development of School and Sport Burnout in Adolescent Student‐Athletes: A Longitudinal Mixed‐Methods Study.” *Journal of Research on Adolescence*, vol. 30, no. S1, Jan. 2020, pp. 115–33. *DOI.org* , https://doi.org/10.1111/jora.12453.

Westfall, Scott, et al. *The Association Between the Coach-Athlete Relationship and Burnout Among High School Coaches*.

Wilczyńska, Dominika, et al. “Burnout and Mental Interventions among Youth Athletes: A Systematic Review and Meta-Analysis of the Studies.” *International Journal of Environmental Research and Public Health*, vol. 19, no. 17, Aug. 2022, p. 10662. *DOI.org* , https://doi.org/10.3390/ijerph191710662.

Applications of Machine Learning to Animal Species Classification By Brady Wan

Abstract

In this project, we use supervised machine learning methods to predict outcomes. In this project, we consider a dataset that shows how machine learning can be applied to animal species classification.

1. Introduction

Machine Learning is the combination of computer science and statistics. Machine learning uses data sets to create models and those models are used to make predictions on new data with unknown outcomes. We refer the readers to the books [1, 2, 3, 4] for more details.

The problem we consider in this article is a supervised learning problem, which is a type of machine learning problem that uses data sets to make models and predict outcomes.

This specific problem deals with species classification of penguins. Since there are more than 2 types of categories that a penguin could be predicted to be, this is a multi classification problem. The data set given is data on various penguins that are either Adelie, Chinnstrapm or Gentoo species penguins. Our data set contains 343 different penguins. It contains data on the penguins such as their species, the island they inhabit, their culmen length in millimeters, and more as shown. This type of problem is known as supervised machine learning. This is a technique in machine learning that uses given datasets to make algorithms or models to most accurately predict outcomes. In this case of supervised machine learning, we will try to predict a penguin species between Adelie, Chinstrap, And Gentoo penguins. The species of the penguin is what we are trying to determine. This is the label. The other pieces of data that the model will use to try to predict the label are known as the features. In this dataset, every penguin in this data set has 1 label and 6 features. Using the data to make a model to predict is called training the model. We train this model with this data set. Since the species is already known, it can be used to first test the model and make the model until it's able to accurately predict species of penguin using the features reliably.

	species	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	MALE
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	FEMALE
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	FEMALE

The Species shown on the very left is the label, while the other 6 columns are the features.

2. Cleaning Up Data

Before being able to train the model, we must first 'clean up' the data. For example, penguins with missing data we must remove from the data set, or categories that are categorical (non numerical), we must one hot encode. One hot encoding is the way for the computer to

understand data that isn't numerical. We do this with 1's and 0's. For example, the column 'sex' has male, female, and '.'. We first remove the penguin with the '.' as the sex. We then simply replace male with 1 and female with 0 to have only numerical values for the column. For one hot encoding the species and islands inhabited is a little different because there are three different options instead of two. In this case, There are 3 different possible labels: the 3 different types of penguin species. In our given dataset, each species of penguin gets a 0 or 1. Whichever species the penguin actually is gets a 1, and the other two get a 0. This is how we one hot encoded multi classification data. We will create 3 new columns: one column for each species. Then, for example, a penguin in our dataset that is part of the Adelie species will have a 1 under the Adelie column and 0's in the other two columns. Lastly, we must also remove any penguins with missing data. We filter penguins with missing data and delete them from the data set.

	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	sex	Torgersen	Biscoe	Dream	Adelie	Chinstrap	Gentoo
0	39.1	18.7	181.0	3750.0	0	1	0	0	1	0	0
1	39.5	17.4	186.0	3800.0	1	1	0	0	1	0	0
2	40.3	18.0	195.0	3250.0	1	1	0	0	1	0	0
4	36.7	19.3	193.0	3450.0	1	1	0	0	1	0	0
5	39.3	20.6	190.0	3650.0	0	1	0	0	1	0	0

We get rid of the species column and replace it with 3 columns, each with one of the 3 penguin species provided in the dataset. Penguins 0-5 are all part of the Adelie Species.

3. Multi- Classification Problem

As mentioned before, our problem is a multi classification problem. A penguin can be part of the Adelie, Chinstrap, or Gentoo species. In a multi classification model, the model takes the features of the penguin as the input (island, culmen length, flipper length, etc). Then, the output will be a series of 3 numbers, since there are 3 possible classifications. For example, an output for a predicted penguin's species could be [0.2, 0.6, 0.2], with each decimal corresponding to each of the 3 species: Adelie, Chinstrap, Gentoo. These decimals can be seen as probabilities. There is a 0.2 or 20 percent chance the penguin species is Adelie based on the input(features) given, a 0.6 or 60 percent chance the penguin species is Chinstrap based on the input given, and a 0.2 or 20 percent chance the penguin species is Gentoo based on the input given. Since this particular penguin has the highest probability to be a Chinstrap species based on the data, the output will turn into [0,1,0]. This would translate into Chinstrap being the predicted species for that penguin.

4. Softmax

Softmax is the function we use for a multi classification problem. This function is this

$$s(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$$

$S(x_i)$ represents the probability of class x_i. For example, the probability of a penguin being part of the Adelie species. Each species type in our dataset is a class. The most important property for this function is that the function will produce a number between 0 and 1, (as the function is producing probabilities) and all the classes values combined will equal 1. For example, this function will produce a probability of a penguin being an Adelie, a probability of a penguin being a Chinstrap, and a probability of a penguin being a Gentoo species. These 3 probabilities will sum to equal 1. The softmax function in this case will take in an input of 3 classes, ei. [12, 30, 9]. Each number represents one of the species. So 12 represents Adelie, 30 represents Chinstrap, and 9 represents Gentoo. These numbers are just the numbers that were calculated through the machine learning model developed. The softmax function is applied at the very end, where the machine will then produce probabilities. Take for example the example input for a penguin x. we have: [3, 5, 1]. The output in terms of variables would be [x_1, x_2, x_3]. We would replace the x_1 with the 3, x_2 with the 5, and x_3 with the 1. So, we first take $e^3/(e^3+e^5+e^1)$. We replaced the x_i in the original formula with x_1. The denominator is just the sum of each input value when it's raised from the e power. When we do this calculation, we get 0.117. This means the model predicts that there is a 11.7 percent chance that penguin x is an Adelie species penguin. Then, we move on to x_2. We make the numerator e^5 and the denominator stays the same since it's always the sum of the three numbers in the input. We get $e^5/(e^3+e^5+e^1)$ and that equals around 0.867. Therefore, the model predicts that there is a 86.7 percent chance that penguin x is a Chinstrap species. Lastly, for the calculations for $e^1/(e^3+e^5+e^1)$, we get 0.016. Therefore, the model predicts that there is a 1.6 percent chance that penguin x is a Gentoo species. In this scenario, the model would predict that penguin x is part of the Chinstrap species. Also notice all 3 probabilities add up to 1.

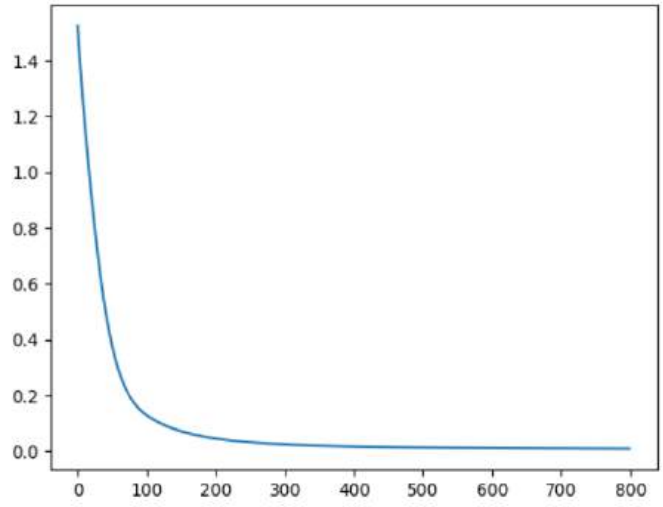
5. How to determine accuracy of the model

When we train a model, there is something called overfitting. This is where the model adopts too much to the data given and tries to predict the randomness that comes with anything. This is bad because while it may have high accuracy to the dataset we gave it, it wouldn't be as accurate to new data because it's still trying to use the randomness from the previous dataset to make these predictions. This would make the model unusable because we need models that can be accurate in new data, not just data it was originally given. To check for overfitting, we split our data into the training set and validation set. The training set is used to train the model, and the validation set is used to test the trained model on new data. Essentially, when we check accuracy for these sets, we want to see good accuracy in both the training and validation set. In order to train this model, we have to set a number of epochs. 1 epoch represents 1 iteration the model goes through the data(in our case, the 343 penguins). After each iteration, the model will

pick up which labels it predicted correctly, which ones it didn't, and attempt to fix the model for better accuracy. Our job is to see at how many epochs it takes for the model's accuracy to be the greatest without overfitting.

	accuracy	loss	val_accuracy	val_loss
0	0.485944	1.112550	0.440476	1.137149
1	0.570281	1.059686	0.488095	1.095992
2	0.582329	1.020647	0.547619	1.060276
3	0.586345	0.985634	0.559524	1.026937
4	0.598394	0.952739	0.583333	0.995262
...	...	...	...	...
795	1.000000	0.007669	1.000000	0.003489
796	1.000000	0.007661	1.000000	0.003485
797	1.000000	0.007679	1.000000	0.003481
798	1.000000	0.007652	1.000000	0.003479
799	1.000000	0.007653	1.000000	0.003473

The numbers in bold on the far left are the number of epochs. As you can see, as the number of epochs goes up, the accuracy for both the training set and validation set are basically 100 percent. Notice when the machine only iterated and tested the dataset once or twice, the accuracy was pretty low. Additionally, the graph below shows the epochs vs loss. decreases as the epoch



As you can see, the loss increases, but then plateaus. This is how we know when to stop adding more epochs, as the loss is not getting any smaller after a certain number of epochs.

6. Conclusion

Machine learning is a powerful field of computer science that can be applied to various different fields. In this article, we show how machine learning can be applied to animal species classification.

Works Cited

[1] Andriy Burkov. The hundred-page machine learning book, volume 1. Andriy Burkov Canada, 2019.

[2] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. The elements of statistical learning, volume 1. Springer series in statistics New York, 2001.

[3] Tom M Mitchell et al. Machine learning. 1997.

[4] Toby Segaran. Programming collective intelligence: building smart web 2.0 applications. ” O’Reilly Media, Inc.”, 2007.

How the Lack of Corporate Accountability by Shell and Other Multinational Oil Companies has Fueled Vast Inequality in Nigeria, Despite its Oil Wealth, with Government Corruption and the Failure of International Law Allowing Neo-colonial Exploitation of Resources By Tammy Oyebanji

Abstract

Despite Nigeria's oil wealth, the country faces extreme inequality. This is because of the exploitation of its resources by multinational companies like Shell. However, regardless of the billions that multinational oil companies and the government have taken from the Niger Delta and other small villages in Nigeria, they do little to assist the local population, leaving many Nigerians in poverty. The government fails to regulate these companies or hold them accountable due to corruption, as politicians benefit and profit from the situation. This makes it easier for multinational companies to continue their harmful practices. This exploitation can be seen as a form of neocolonialism, where foreign companies and the Nigerian government extract Nigeria's resources, while the local population remains impoverished and does not share in the economic prosperity or wealth generated by these multinationals. International laws are also weak and do not do enough to stop these companies from taking advantage of Nigeria.

This paper explores the intersection of corporate greed, government corruption and failure of international law, showing how these factors contribute to Nigeria's inequality. By framing these issues through the lens of neo-colonialism, this review shows how systemic structures allow the exploitation of Nigeria to continue. It fills a gap by connecting historical exploitation with modern practices and also bridging environmental, economic and legal perspectives.

Neo-colonialism: A Theoretical Framework for Corporate Exploitation

The ongoing exploitation by Shell and other multinational oil companies in Nigeria over the years represents neo-colonialism. "Neo-colonialism is the worst form of imperialism. For those who practise it, it means power without responsibility and for those who suffer from it, it means exploitation without redress" (Tiger & Nkrumah, 1966). This exploitation is not just economic; it is a continuation of colonial legacy, where foreign powers hold sway over a nation's resources, like oil, leaving the local population to bear the burden of wealth extraction.

Oil, in this context, becomes more than a resource; it is a symbol of corruption and the lasting effects of colonial rule in Nigeria (Best Documentary, 2024). The revenue from oil only benefits 1 to 5% of the Nigerian population (Best Documentary, 2024). Nigeria's economy is heavily reliant on oil, which allows multinational oil companies to exploit the situation. As a result, the main beneficiaries are the oil companies and the government's corrupt officials. This situation reflects the colonialist nature of these oil companies, which exploit the country's resources without giving anything back to the local communities. "The result of neo-colonialism is that foreign capital is used for the exploitation rather than for the development of the less developed parts of the world. Investment under neo-colonialism increases rather than decreases

the gap between the rich and the poor countries of the world" (Tiger & Nkrumah, 1966). "In old-fashioned colonialism, the imperial power had at least to explain and justify at home the actions it was taking abroad. In the colony, those who served the ruling imperial power could at least look to its protection against any violent move by their opponent. With neo-colonialism, neither is the case" (Tiger & Nkrumah, 1966).

The lack of international laws to prosecute these corporations allows neo-colonialism to persist, which is essentially just colonialism with a new facade. There is limited corporate accountability, as demonstrated by the continuation of exploitative practices reminiscent of colonial rule. The fundamental injustice lies in the billions of dollars that have been taken from Nigeria with little to no development in the communities around the Niger Delta where the majority of oil revenue is from. Colonialism involves the extraction of resources and exploitation, and oil companies have consistently done this across Africa and international law does not hold these companies accountable. As noted, "Only if cases of abuse reach the international press does it appear that there is even a chance that the oil companies may publicly register concern with the authorities. And even then, it is more likely that any concern voiced will be private, and thus remain unproven" (Manby, 1999). It is perplexing that companies can so easily adopt colonial attitudes and practices without facing consequences. "In retrospect, it was perhaps unrealistic to hope that a political system largely inherited from the British colonialists under the direction of a political class adapted to that system could become anything but corrupt" (Cayford, 1996).

Shell and the other multinational oil companies have taken advantage of this unique situation. The modern dynamic of unchecked corporate power mirrors the colonial monopolies that extracted resources without community input or benefit. Shell collaborates with the government to secure oil profits while neglecting its social and environmental responsibilities.

The Historical Context of Oil Exploration in Nigeria

The discovery of oil in Nigeria dates as far back as 1937 when Shell D'Arcy acquired a license for petroleum exploration for the whole of Nigeria. It was not until 1956, under a joint venture with the British government, the multinational Royal Dutch/Shell discovered the first 'commercially viable' Nigerian oil field. The discovery marked the beginning of Shell's tumultuous relationship with this small region known as Ogoniland within the Niger Delta" (Cayford, 1996). The growing importance of oil into the economy of Nigeria was shown by the first oil export in 1958 at a rate of 5100 barrels a day.

After independence, there was little regulation of the industry by the government and this resulted in increased production levels in the 1960s and 1970s. Nigeria joined the Organization of the Petroleum Exporting Countries (OPEC) in 1971 and set up the Nigerian National Oil Corporation (NNOC), renamed the Nigerian National Petroleum Corporation (NNPC), to manage oil production. In spite of booms from oil, the government continually became dependent on their revenues, producing a "resource curse"; such a case occurs where other industries are wholly neglected.

While the government and oil multinationals were making profits, communities in the Niger Delta faced serious environmental problems due to oil extraction. There were many oil spills, gas flaring, and land damage, harming local ecosystems and ruining traditional jobs like farming and fishing. Political instability further exacerbated these issues. During the military regimes of the 1970s and 1980s, oil money fed corruption and disparity between the rich and the poor. Instead of investing in public infrastructure or community support, government officials lined their pockets, often with the cooperation of foreign companies like Shell and Chevron. It established systemic inequality in ways because the oil wealth was not shared in the communities where it came from.

These issues were also an extension of the effects of colonial exploitation. During colonial rule, resource extraction mainly served British interests, leaving Nigeria with little infrastructure for fair development. After gaining independence in 1960, Nigeria continued these exploitative practices, creating a system where foreign companies dominated oil production while ignoring environmental and social concerns.

Today, the Niger Delta is possibly one of the most polluted places in the world because of the frequent oil spills and flaring of gases (Obi, 2023). Despite generating billions of naira in oil revenue, the area has high levels of unemployment, poor infrastructure, and political neglect. The history of oil exploration in Nigeria shows the complex issues of foreign exploitation, government corruption, and economic inequality that still happen to this day.

The Devastating Environmental and Economic Consequences of Oil Extraction

With the heavy presence of the oil exploration activities of the multinational oil corporations, the livelihood of the Nigerians has been severely destroyed (DW News, 2012). Shell and other multinational oil companies have ruined the environment and there is no corporate accountability for their actions. A billion-dollar industry has given nothing back to the communities it is harming. Oil spills have devastated the livelihoods of many people. In regions where oil companies operate, traditional means of earning a living have been severely impacted (VICE News, 2018).

Agriculture, which is a major source of income and employment in rural areas of Nigeria, has suffered greatly due to the activities of Shell and other multinational companies (DW News, 2012). The oil spills have polluted the seas, creating a lack of clean drinking water and a decline in fish populations (Bloomberg Originals, 2016). This has further limited income opportunities for local communities.

As a result, some individuals resort to illegal activities to survive, leading to a significant decrease in life expectancy (Best Documentary, 2024). Many people are now forced to risk their lives by tapping into oil pipelines for a monthly wage of around £100 (Best Documentary, 2024). This unemployment also links back to Shell and the other multinational oil companies because “most of those employed by the oil industry are not from the oil-producing communities. Shell's Nigerian operation, for example, employs only some 5,000 people, and due to poor education in

the Delta most of them are from other parts of Nigeria, especially the Yoruba and Igbo ethnic group" (Manby, 1999).

The situation is dire, with locals reporting rashes and various skin problems due to the heavily polluted environment surrounding them (Al Jazeera English, 2023). Many residents feel they are being treated unjustly, almost like animals (DW News, 2012). The issues in Ogoniland, which accounts for roughly 90 percent of Nigeria's oil production, extend beyond oil spills. Air pollution and soot from continually burning gas flares, often located next to villages, pose significant health risks (Best Documentary, 2024). Additionally, poorly designed causeways and canals used by the oil industry disrupt the hydrology of the seasonally flooded freshwater swamps and the brackish waters of the mangrove forests, leading to the death of crops, devastation of fishing grounds, and contamination of drinking water supplies (Al Jazeera English, 2023). The placement of pipelines and other oil facilities has also severely impacted the land, consuming vast amounts of valuable farmland. The Ogoni people contend that they are poorly compensated for this loss and are often not consulted about these developments (Al Jazeera English, 2023). Despite the oil industry's considerable contributions to Nigeria's economy, the benefits for the Ogoni community have been either minimal or negative. After two decades of oil exploration, Shell has built only one road in the region (Cayford, 1996).

Neither the federal nor state governments have shown any interest in improving the area or the quality of life for its residents (Cayford, 1996). Oil production and the development it supposedly fosters have caused significant environmental damage in the Niger Delta without delivering proper benefits to the local population. While some individuals have become wealthy from oil money, the majority of people living in oil-producing regions remain impoverished. This inequality in the distribution of oil wealth has intensified conflicts both within and between communities (Best Documentary, 2024). It is astonishing that, despite multinational oil companies generating billions of dollars in profit, there has been no meaningful development in these areas. The slow progress can only be attributed to "the long neglect of the region by the Federal Governments of Nigeria and the nonchalant attitude of the companies operating in the area" (Onweazu, 2012). "Development involves material improvement no doubt, but it also demands that such material benefits should be enjoyed by all and not by a privileged few" (Onweazu, 2012). "In short, capitalism must be made responsible for its wider environmental and social impacts" (Maiangwa & Agbiboa, 2013).

The Role and Challenges of the Amnesty Programme

In Nigeria, many citizens feel helpless because of the inadequate support from their government and international law. The sense of insecurity often makes people form collective groups in an effort to protect themselves from exploitation going on in their country. Many groups were formed because of the problem of oil exploitation.

The situation escalated when "the Nigerian government lost a staggering USD 23.7 billion in oil revenue due to The Movement for the Emancipation of the Niger Delta (MEND) attacks" (Maiangwa & Agbiboa, 2013). MEND emerged as a militant group in response to the

exploitation and environmental destruction caused by multinational oil companies and government corruption. MEND was founded in the early 2000s. Their mission was to fight for greater control over the Niger Delta's oil wealth and demand improved living conditions for the marginalised communities most affected by resource extraction.

The extreme financial loss forced the Nigerian government to address the issue of oil in Nigeria, showing the extent of government corruption and their indifference to the suffering of communities in the Niger Delta. The government only took action when their interests were at stake. In an attempt to address and cover up this situation, the Nigerian government made the Amnesty Programme.

"The Amnesty Programme was announced by President Yar'Adua on 25 June 2009. The terms stated that militants who freely surrendered their arms within the 60-day amnesty period (6 August 2009 to 4 October 2009) would not be prosecuted for the crimes that they had committed while disrupting the Nigerian oil industry" (Maiangwa & Agbiboa, 2013). This was a significant moment in Nigeria because it was the first time the government directly engaged with groups like the MEND, who sought greater autonomy, justice, and economic empowerment for their region.

In contrast to MEND, the Movement for the Survival of the Ogoni People (MOSOP) took a non-violent stance, focusing on issues such as environmental degradation, revenue sharing, and corporate accountability (Cayford, 1996). Despite their peaceful approach, MOSOP's efforts largely went unacknowledged. They also got severe backlash from the government, which responded with brutality rather than dialogue (Manby, 1999). The government's violent tactics included banning public gatherings, breaking up peaceful protests, demolishing entire communities, and even killing demonstrators. This was topped in the unlawful execution of Ken Saro-Wiwa, the leader of MOSOP, along with eight other Ogoni leaders in 1995, intended as a warning to future activists (Cayford, 1996).

This contrast between groups shows the government's corrupt nature; while they resorted to violence against MOSOP's peaceful protests, they had to take action with MEND because of the significant financial loss and global scrutiny the situation was attracting. MEND's armed resistance included kidnapping oil workers, attacking oil facilities and sabotaging pipelines (Maiangwa & Agbiboa, 2013). The reality was that the threat of losing billions forced the government to reconsider its approach to the ongoing crisis in the Niger Delta.

The militants wanted a solution. Many militants turned themselves in even though key militant groups like MEND viewed the amnesty with suspicion because it created less room for dialogue and did not address the core issues that had given rise to the struggles in the first place (Maiangwa & Agbiboa, 2013). It was merely an excuse from the government; they did not want to address the issues going on in their country, providing no solutions to the problems. They simply wanted to stop what directly affected them.

"Specifically, cash payouts to armed militants and proposals to give oil-bearing communities a 10 per cent stake in state oil revenues fail to seriously address the core underlying issues (e.g. government corruption, the political sponsorship of violence, environmental

degradation by oil MNCs that continue to fuel hostilities and resistance in the Niger Delta. According to Omeje, what prompted the amnesty proposal was not the environmental tragedy unfolding in the Niger Delta but rather the urgent need to stem the tide of crippling MEND attacks on oil facilities in Nigeria. These negatively affected the country's oil productivity and the profits of oil MNCs in the region, especially Shell's. In short, the Nigerian state's prime concern in the management of the conflict has always been to maximise and protect oil revenues" (Maiangwa & Agbiboa, 2013).

The Nigerian government failed to implement the programme effectively. There were allegations that funds meant for the reintegration of former militants were embezzled by corrupt officials, which hindered the programme's success. Some militants reported that they did not receive the financial support or training they had been promised (VICE News, 2018). Also, the lack of international help was a significant issue. This is because there was minimal pressure from international organisations or the global community to ensure that the government stuck to its commitments. The government's actions frequently contradicted international standards. Despite the amnesty programme, the Niger Delta region continued to suffer from poverty, environmental degradation, and underdevelopment (VICE News, 2018).

Corporate Social Responsibility: Promises, Failures, and Community Impact

The response of oil multinational companies to the socio-economic and environmental issues in the Niger Delta can be seen through micro and macro corporate social responsibility (CSR). CSR is a company's commitment to positively impacting society and the environment, beyond making profits.

Micro CSR initiatives, such as community development projects, often fall short of addressing deeper systemic problems like poverty and environmental degradation caused by oil exploration. These efforts, including building schools or giving limited local employment, are seen as superficial (Onweazu, 2012). On the other hand, macro CSR practices can worsen issues like corruption and inequality because of the significant difference in oil revenues. The disconnect between corporate promises and actual practices reveals a lack of commitment to environmental sustainability and community welfare. As a result, the Niger Delta continues to face environmental destruction, poor living conditions, and strained relations with oil companies, showing the inadequacies of CSR efforts (Onweazu, 2012).

Interestingly, while host communities show dissatisfaction with these efforts, corporate employees often believe they are making a positive impact. This disconnect is evident in the following table, which shows employees' attitudes toward their companies' CSR initiatives.

Fig 1: To what extent have the Sustainable Development Programmes of Multinational Oil Corporations conform to the standard of corporate integrity, commitment and moral maturity?

		SD	D	A	SA	Mean	S.D.
1	Good corporate citizenship is absolutely necessary for the sustenance of a healthy operating environment for multinational oil corporations	2 2.0%	6 6.0%	13 13.0%	79 79.0%	3.69	.68
2	Information management technique of the corporations is important for the development of corporate integrity competence with host communities	4 4.0%	1 1.0%	19 19.0%	76 76.0%	3.67	.70
3	If the multinational oil corporations build the intervention programmes around corporate integrity this will improve their relationship with the host communities	2 2.0%	4 4.0%	41 41.0%	53 53.0%	3.45	.67
4	Corporate integrity capacity ensures that multinational oil corporations view seriously its social responsibility in the area of environmental protection	4 4.0%	16 16.0%	24 24.0%	56 56.0%	3.32	.89
5	The corporate social responsibility response of the multinational oil corporations to the host communities is influenced by self discipline and self-control which is morally defined	9 9.0%	8 8.0%	34 34.0%	49 49.0%	3.23	.94
6	Corporate integrity which is bounded by morally justifiable principles and values will enhance the oil corporations to improve on the intervention programmes	5 5.0%	17 17.0%	36 36.0%	42 42.0%	3.15	.88
7	Inability of the oil corporations to address the environmental degradation which occur as a result of oil exploration has affected corporate integrity, morality and commitment	49 49.0%	31 31.0%	15 15.0%	5 5.0%	1.76	.89

(Onweazu, 2012)

As shown in the table, most corporate employees believe their companies' CSR initiatives work well. In fact, 79% say that good corporate citizenship sustains a healthy operating environment. However, this positive view is very different from what local communities' experience. Many of them feel unhappy, left out, and suffer from environmental damage.

Conclusion

Although efforts have been taken to address the exploitation of oil in Nigeria, it is evident that much stronger international accountability for these abuses is needed. Also, there needs to be effective measures against government corruption, and a shift to community led development in the ongoing environmental devastation and socioeconomic disparity across the Niger Delta. This is why understanding how neo-colonial exploitation worsens the problems of environmental degradation and economic inequality is relevant. As, it can help reveal actionable suggestions and pathways toward sustainable development, ensuring the wealth accruing from the resources is returned to those most affected. If implemented properly, these measures can pave the way towards a more just and sustainable future in Nigeria. However, The Niger Delta will still remain an example of how corporate greed and weak governance keep inequality and environmental harm in existence.

Works Cited

Al Jazeera English. (2023, April 4). Shell court case: Oil giant accused of pollution in the Niger Delta [Video]. YouTube. https://www.youtube.com/watch?v=6KNLWcg73M4

This source talks about a court case where Shell is being sued for causing pollution in the Niger Delta. It connects to the theme of how companies should be legally accountable for the harm they do to the environment and people.

Best Documentary. (2024, June 27). The Niger Delta, a war for crude oil [Video]. YouTube. https://www.youtube.com/watch?v=St7yWvGTDmo

This source talks about how Shell and other big companies are making the environment worse, violating human rights, and causing inequality in the Niger Delta. It connects to the bigger theme of how companies keep exploiting the region, making things worse for the people there.

Bloomberg Originals. (2016, March 7). Money alone can't fix the Nigerian village ruined by Shell's oil [Video]. YouTube. https://www.youtube.com/watch?v=y2yPii0yz10

This source talks about a settlement given to Bodo residents by Shell for oil spills, but even with the money, they still live in poverty. It connects to the theme of how the compensation from companies like Shell is not enough to fix the long-term damage they've caused.

DW News. (2012, November 12). Nigeria: Oil pollution in the Niger Delta | Global 3000 [Video]. YouTube. https://www.youtube.com/watch?v=3A-tLtqM8YU

This source discusses the effects of oil spills in the Niger Delta, showing how much contamination oil companies have caused. It connects to the need for companies to be held accountable for the harm they cause and the calls from communities for justice.

Maiangwa, B., & Agbiboa, D. E. (2013). Oil Multinational Corporations, Environmental Irresponsibility and Turbulent Peace in the Niger Delta. Africa Spectrum, 48(2), 71-83. https://doi.org/10.1177/000203971304800204

This source talks about the Niger Delta Amnesty Programme, which tries to stop violence and help militants rejoin society. However, it shows how hard it is to solve conflicts that oil companies and the government cause. This connects to the idea that companies and governments are still not being held responsible for their actions.

Manby, B. (1999). The Role and Responsibility of Oil Multinationals in Nigeria. Journal of International Affairs, 53(1), 281–301. http://www.jstor.org/stable/24357796

This source looks at how oil makes problems in the Niger Delta worse, causing violence and making political and economic issues bigger. It shows how armed groups and corruption keep the conflict going. This connects to the bigger idea of how foreign companies and government corruption keep causing regional problems.

Obi, O. (2023, May 19). Oil among the mangrove trees: A portrait of destruction in the Niger Delta, then and now. Harvard International Review. https://hir.harvard.edu/oil-among-the-mangrove-trees-a-portrait-of-destruction-in-the-niger-delta-then-and-now/ This source provides a vivid account of the environmental and social devastation caused by oil extraction in the Niger Delta region of Nigeria. This

connects to the theme of how oil companies exploit the region, making things worse for the people living there.

Onweazu, O. O. (2012). Multinational oil corporations corporate integrity ethics and sustainable development in Niger Delta, Nigeria. Journal of Sustainable Development, 5(10). https://doi.org/10.5539/jsd.v5n10p114

This source shows how the oil in the Niger Delta hasn't helped local communities, as oil companies focus on profits and ignore environmental and social problems. It connects to the theme of how companies exploit the region, making things worse for the people living there.

Steven Cayford. (1996). The Ogoni Uprising: Oil, Human Rights, and a Democratic Alternative in Nigeria. Africa Today, 43(2), 183–197. http://www.jstor.org/stable/4187095

This source talks about the damage Shell caused in Ogoni land, which has hurt the environment and local communities. It connects to how oil companies are not being held accountable, and how their actions cause long-term damage and inequality.

Tiger, L., & Nkrumah, K. (1966). Neo-colonialism. the last stage of Imperialism. International Journal, 22(1), 161. https://doi.org/10.2307/40199801

This source explains how neocolonialism sustains foreign control over African nations through economic and political means, despite their independence.

VICE News. (2018, March 22). The battle raging in Nigeria over control of oil [Video]. YouTube. https://www.youtube.com/watch?v=vAgw_Zyznx0

This source talks about the pollution, corruption, and poverty in the Niger Delta, showing that the people haven't benefited from the oil wealth. It connects to the theme of exploitation, where oil companies take advantage of the region while leaving the people behind.

Enhancing Diversity and Equity in Dermatology: Improving Representation and Diagnostic Accuracy for Equitable Care for All Skin Tones By Janani Muniswamy

Abstract

In the United States, the medical field of dermatology has a staggering lack of diversity in medical training and representation among doctors. This article addresses methods to improve and redesign diagnostic tools, educational curricula, and cultural competence to enhance diversity in dermatology to ensure equitable care for all skin tones(in fact insert data specific to dermatology. The field of dermatology remains mainly focused on lighter skin tones rather than skin of color where skin of color is accurately diagnosed at below 50%, and less than 20% of images depict skin conditions on skin of color. Through careful examination of dermatology training programs, diagnostic tools, and current cultural competency statistics, I propose steps forward for the field of dermatology to increase diversity and equity. The overall findings of this paper conclude that there are clear methods to improve diversity in dermatology, including redesigning dermatology curricula, utilizing improved and advanced diagnostic tools, and diversifying the dermatology workforce. Given that there is a majority-minority population expected by 2044, it is imperative for steps and changes to be made to the field of dermatology to ensure equitable care for patients of all racial and ethnic backgrounds.

Introduction

Dermatology holds its rank as one of the least diverse medical specialties (Williams and Shinkai). Skin conditions are among the most common in primary care, but there are discrepancies in diagnostic accuracy on skin of color. The diagnostic accuracy of cutaneous/subcutaneous pathology in dark skin is, on average, only 44.3% (Perlman et al.). This is drawn back to its variation in presentation between white skin and skin of color (Bellicoso et al.). *"To this day, there's inappropriate treatment of skin of color. Then Black people get the sense that, well they don't really know my hair, they don't really know my skin,"* says Chesahna Kindred, MD, a board-certified dermatologist in Columbia (*Inside the DEI Battle in Dermatology | Allure*). In fact, a survey of 140 dermatologists and dermatology trainees found that approximately one-third to one-half of academic dermatologists and residents reported inadequate training in skin conditions affected in patients of color (Narla et al.). Additionally, from 1995 to 2005, only 2% of the total teaching events at the annual meetings of the American Academy of Dermatology were devoted to skin of color. In a review of popular dermatology textbooks, at most, 19% of pictures were of patients with skin of color. Also, less than 7% of all images of skin diseases were skin of color (Bellicoso et al.). *"The biggest failure would be that the public loses faith in their doctors,"* says Dr. Mariwalla, who also believes the lack of DEI (diversity, equity, and inclusion) will negatively impact patient care (*Inside the DEI Battle in Dermatology | Allure*). Given the issues presented, there is bias in diagnostic tools, causing misdiagnosis, and a lack of inclusivity in many of the current dermatology curriculums. In 2044, a majority-minority population is expected only highlighting the call for greater action (Williams

and Shinkai). It is imperative to take steps to improve diversity and cultural competency in dermatology.

To aid the diverse palette of skin tones, the dermatology field must first embrace the vibrant colors it has within. Dermatologists specialize in the study of the skin, also treating and diagnosing conditions related to the hair, nails, and mucous membranes (*Dermatologists: What Do They Do, Qualifications, and Procedures*). The education and training path of a dermatologist takes 12 or more years to complete (*How to Become a Dermatologist - Guide to Dermatology and the Path to Becoming Dermatologists*). Several critical steps are required to become a dermatologist, beginning with earning a bachelor's degree. Next, students must pass the Medical College Admissions Test (MCAT), and attend medical school to learn the basics of practicing medicine. After finishing medical school, a student must complete a four-year residency program. Also, at the end of residency, to become a licensed practitioner, a student must take and pass the last part of the USMLE: which is a requirement to earn a license to practice dermatology. Finally, to be board certified, one must have a medical degree from an accredited medical school, a valid dermatology license, pass the standardized ABOD exam, and complete a fellowship program. This board certification requires a renewal every ten years. However, this extensive process, although rewarding, can be extremely expensive, which can restrain individuals from lower socioeconomic backgrounds who may not have the resources to afford expensive education (Fernandez). Additionally, faculty members of color in academic medical institutions must pay “minority tax” (“Perspective”). “Minority tax” entails the substantial time colored faculty members must devote to service on committees, panels, and boards on diversity, equity, and inclusion. This service is usually uncompensated and can impede career development. This is not a time commitment that white skin faculty members worry about, who spend time on activities that can further advance their careers.

In residency programs, a large focus is placed on clinical dermatology (“Curriculum”). For instance, at the University of Pennsylvania, the curriculum involves significant time in clinical rotations, didactics, and lectures. The topics covered for dermatology students range from general dermatology to dermatopathology. This curriculum is common for most dermatology training programs and is required to become a dermatologist. While the general dermatology curriculum seems broad, it faces criticism due to its lack of inclusivity, primarily through a lack of training in skin conditions on darker skin tones, an insufficient emphasis on cultural competence, and a deficiency of representation within the dermatological field (Narla et al.). Together, these factors have led to the manifestation of racism in the dermatology curriculum. The issue is further exacerbated by socioeconomic factors and the underrepresentation of minority groups in the dermatological workforce. These factors can also lead to misdiagnosis, which can cause larger implications. In a study “Prevalence of misdiagnosis of cellulitis: a systematic review and meta-analysis,” Todd S Cutler found that the misdiagnosis of cellulitis is common and can lead to unnecessary hospital admission and antibiotic overuse. This paper will investigate how dermatology training programs could improve

inclusivity both in the curriculum and diagnostic tools and enhance cultural competence to champion equitable care for all skin tones.

Inclusivity in dermatology training programs

In a study called "Evaluating Diversity and Inclusion Content on Graduate Medical Education Websites" conducted by Chapman Wei et. al, certain, but not all, dermatology programs have expanded efforts into increasing inclusivity within dermatology. Utilizing key factors, such programs ensure diverse care for all skin tones to increase student confidence in treating and caring for patients of a wide range of skin tones. Some examples of schools intentionally implementing inclusivity and anti-racism training within their programs include the University of California, San Francisco, University of Miami Miller School of Medicine, Howard University, Harvard University, and the University of Texas Southwestern Medical Center. Within these programs, increased efforts have been put towards specific key factors. The first is diversity in clinical images and case studies. This covers the representation of all skin tones to be used for imaging and research purposes, creating diverse research and literature. Next is dermatoscopy and diagnostic techniques, covering the current tools used in dermatology. dermatoscopy specifically refers to the examination of the skin using skin surface microscopy (mostly used to evaluate pigmented skin lesions) (Sonthalia et al.). In addition to that, cultural competence training encompasses components such as cultural awareness, sensitivity, communication, inclusivity, and more. Similarly, diverse faculty and leadership promote diverse patient care with a broader range of knowledge and skills. Another is community outreach and engagement to ensure feedback is given and acted upon (Bohler et al.). The mentioned initiatives aim to ensure diverse care for all skin tones by using these key factors to increase exposure to a wide variety of patients. These efforts collectively build inclusivity within programs, promoting a diverse dermatology workforce in the future.

With certain efforts to increase inclusivity and diversity within dermatology training programs, studies have backed these initiatives. Studies show that these efforts have positively affected dermatology training. For example, a study, "Evaluation of a skin of color curriculum for dermatology residents," examined the impact of teaching a skin of color curriculum for a week. The curriculum contained one-hour lecture sessions with information specific to the cultural and medical nuances of skin of color. Five lectures were held within a one-week time frame, and the lecturers included two academic physicians with clinical and research focus on skin of color as well as one dermatology resident with an interest in skin of color. The topics covered in the lectures were chosen based on existing gaps in dermatology resident education and the overall relevance of caring for patients of color. It was concluded that 100% of learners felt their ability to care for patients of color improved. They felt that the skin or color curriculum should be an annual component of their dermatology academic curriculum, and they all believed that other dermatology residents would benefit from this curriculum. Another study entitled, "Dermatology resident comfort level treating hair conditions with skin of color," used a survey to understand comfort levels pertaining to hair care knowledge and treatment comfort levels. The

survey prompted residents to report how frequently they treat patients with hair-related conditions (ranging from alopecia areata to dissecting cellulitis) on a scale ranging from 1 to 6 respectively, (never or 1 time per year, once every 6–12 months, once a month to once every 6 months, once a week to once a month, every day to once a week, or multiple times per day). Residents were also asked to rank their comfort level from 1 to 5 respectively, (extremely uncomfortable, somewhat uncomfortable, neither comfortable nor uncomfortable, somewhat comfortable, or extremely comfortable) in recognizing and treating each of these conditions. Using the same scale, residents described their comfort level with understanding the basic hair morphology of SOC patients and their comfort in counseling individuals with skin of color on proper hair practices, best hair products, and healthy hair regimens. The results showed dermatology residents were relatively comfortable with common conditions such as androgenetic alopecia and alopecia areata, but uncomfortable with creating healthy hair regimens and discussing natural hair care products. Specifically, residents self-identified their confidence level, which was significantly impacted when caring for patients with skin of color. This discomfort is concerning because it can lead to negligent patient care and health disparities, highlighting the need for comprehensive training in managing hair and skin conditions in diverse populations to ensure all patients receive equitable and effective treatment. Increasing training and ensuring specific care for skin of color is a regular part of the curriculum can help address these gaps and improve overall patient care.

Although efforts have been made, changes and evaluation are still necessary to improve diversifying education content, increasing representation, enhancing mentorship programs, cultural competency training, and community engagement. However, implementing an updated dermatology curriculum is a complex and time‑consuming task. Additionally, a lack of funding and inadequate infrastructure are significant limitations to developing an updated dermatology curriculum (Khalil et al.). Still, for an inclusive future in healthcare, diverse and inclusive education must be met first, emphasizing the importance and the profound impact inclusivity in dermatology training programs can bring.

Bias in Dermatology Diagnostic Tools

Currently, some widely used diagnostic tools in the dermatologic field may account for the lack of inclusivity in dermatology (Sharma and Patel). Specifically, the Fitzpatrick scale has received large amounts of criticism due to its inability to accurately assess racial, ethnic, or phenotypic features. Thomas B. Fitzpatrick developed the Fitzpatrick skin phototypes (FST) in 1972 to classify an individual's tendency to burn or tan when exposed to sunlight (Sharma and Patel). The classifications are commonly used to estimate dosages of ultraviolet B (UVB), which are the invisible rays that come from the sun that cause sunburn, skin darkening and thickening on the outermost layer of skin, and psoralen with ultraviolet A (PUVA), which is a type of photodynamic therapy used to treat skin conditions such as psoriasis, vitiligo, and laser treatments (*Definition of Psoralen and Ultraviolet A Therapy - NCI Dictionary of Cancer Terms - NCI*). Additionally, it is commonly used to predict skin cancer risk. Originally, FST categorized

skin into four types (I-IV), and the scale was later updated to include non-white patients by adding two additional categories: 'brown' (type V) and 'black' (type VI) (Okoji et al.). Now, FST is the standard skin classification system and is often incorrectly used as a proxy for race/ethnicity, despite the bias as a scale centered on light skin.

It is important to emphasize that the FST was intended to classify sun sensitivity, and not to act as a proxy for racial or phenotypic features (Sharma and Patel). The classification of skin color– white, brown, black – was meant to denote complexion rather than a self-identification of ethnic origin. For example, a cross-sectional survey of an ethnically diverse population showed that self-reported pigmentary phenotypes and race were significant but incomplete predictors of Fitzpatrick skin types. Understanding these limitations is crucial for the proper application of the FST in the future.

The FST relies on self-reporting, which holds multiple limitations (Sharma and Patel). The six categories in the FST are designed to include 'brown' and 'black' skin; however, people with mixed heritage or races who identify as black often self-report in a variety of categories, ranging from type IV to type VI. Labeling one's skin as burning or tanning lacks reliability because these words may mean different things to different people. For instance, self-reported FSTs have correlated poorly with sunburn risk as well as physician-reported FSTs. The functionality of Fitzpatrick skin typing in ethnic skin is also under question. For example, Japanese women often self-identify as FST type II which represents fair skin that always burns and tans with difficulty, but Asian skin generally is considered to be non-white. Fitzpatrick himself has also acknowledged that *race* and *ethnicity* are cultural and political terms with no scientific basis (Ware et al.). Given this information, due to the FST's reliance on self-reporting, a discrepancy is present in regards to consistently associating certain phenotype categories to specific skin tones. This only highlights the significant bias and limitations associated with one of the most common skin classification systems, the FST (*Modified Fitzpatrick Scale-Skin Color and Reactivity - JDDonline - Journal of Drugs in Dermatology*). This only emphasizes the call to action for an objective classification system.

Dermatology must seek an objective classification system, and with the rise of artificial intelligence, a technology-based solution may be possible (Okoji et al.). There are, however, other classification systems used around the world. The Kawada skin classification system (1986) is used by the Japanese and was based on personal history of sun reactivity, similar to how the FST classifies skin (Kawada). Although useful, the Kawada system is only limited to Japanese individuals. Another skin classification system is the Glogau scale (1994), which classifies the degree of photoaging and categorizes the amount of wrinkling on the skin based on ultraviolet light exposure ("Glogau Wrinkle Scale"). However, the Glogau scale does not address signs of photoaging in mixed ethnic-racial skin types (Roberts). Additionally, the Goldman world classification was developed to examine various skin color responses to burning, tanning, and post-inflammatory pigmentation (Goldman). The Fanous classification system is used for laser resurfacing, chemical peels, and dermabrasion (Fanous). This system is based on racial-genetic origins and describes six subraces. Finally, the Willis and Earles scale is a proposed classification

system for people of African descent and classifies skin color, reaction to UV light, and association of pigmentary disorders (Willis and Earles). While there is a long list of classification scales used globally, the Fitzpatrick scale remains the standard, despite its numerous limitations. Even with the other scales globally, most seem to possess alarming limitations causing bias within these diagnostic tools (Roberts). Various skin classification systems exist worldwide, each with its own focus and limitations.

However, with the rise of artificial intelligence, the possibilities for new classification systems that promote inclusivity are endless. For example, the Diverse Dermatology Images dataset (DDI), is an image data set created to address diverse skin tones, and uncommon diseases found on all skin tones (*Disparities in Dermatology AI Performance on a Diverse, Curated Clinical Image Set | Science Advances*). The DDI is said to be the first publicly available, expertly curated, and pathologically confirmed image dataset with diverse skin tones. Additionally, the Roberts Skin Type Classification System is said to be a classification system that aims to be a more comprehensive skin-type classification system to meet the needs of inclusivity as the presence and diversity of ethnic skin color grows (*Embracing Diversity: New Skin Type System Addresses Procedures for Treating Skin of Color*). The Roberts Skin Type Classification System is designed to help dermatologists predict a patient's response to dermatologic procedures

Cultural Competency in Dermatology

Despite the strides toward equity in health care, racism is still present (Narla et al.). Health inequity is manifested through social, economic, and environmental disparities in communities of color. Many factors contribute to the lack of diversity in the healthcare field including, decision-makers in public health research, policymakers, medical educators, officials, hospital administrators, insurance and pharmaceutical executives, as well as healthcare providers, play a critical role in creating and maintaining health inequalities. The AAMC Diversity in Medicine: Facts and Figures 2019 report demonstrates that among active physicians, 56.2 percent identified as White, 17.1 percent as Asian, 5.8 percent as Hispanic, 5.0 percent as Black or African American, with 13.7 percent Unknown, making it the largest subgroup after White and Asian. The healthcare provider workforce does not reflect the population it serves in terms of racial and ethnic background (Narla et al.). Indeed, 75.3 percent identify as white alone, 13.7 percent as Black or African American alone, and 6.4 percent as Asian alone (*U.S. Census Bureau QuickFacts*). To serve the needs of a diverse population, the healthcare system must take measures to improve cultural competence, as well as racial and ethnic diversity. Cultural competence is the ability to collaborate effectively with individuals from different cultures (*Cultural Competence: An Important Skill Set for the 21st Century*). Such competence improves healthcare experiences and outcomes. Measures to improve cultural competence and ethnic diversity will help alleviate healthcare disparities and improve racism that has manifested in the healthcare systems (Nair and Adetayo). According to a survey, more than one-third said that it was very important (4018 [19.2%]) or somewhat important (4038 [18.4%]) that their healthcare

professionals understand or share their culture, whereas more than one-half of the respondents (12 014 [51.5%]) said that sharing one's culture was not important at all (Blewett et al.). Given this, by prioritizing cultural competence, representation can be increased in healthcare, in turn improving diversity, equity, and inclusion in the medical field. Creating a workforce that is a clear reflection of the population it serves will provide equitable care for all fostering an irreplaceable trust between healthcare providers and the diverse population it serves.

Boosting cultural competency can be done in a variety of ways, ranging from hiring more faculty experts on the skin of color, to spreading awareness through panels, presentations, and instruction (Kaundinya and Kundu). Its efficacy must ultimately be measured by student knowledge and learning. Assessments are what notify educators of existing gaps in medical students' knowledge. It's been established that there's a lack of education in the content areas of skin of color. As syllabi are expanded to address these gaps in the curriculum, assessments could help to ensure that appropriate knowledge is gained, and notify educators of the success of proposed interventions. Through this process, it could be more manageable to ensure the success of cultural competency, leading to a more culturally humble medical workforce.

Conclusion & Future Directions

In order to achieve diversity in dermatology, we must implement specific changes within inclusivity in dermatology training, redesigning current diagnostic tools to decrease bias, and emphasizing the importance of cultural competence in dermatology. Redesigning and changing what is done now are crucial to creating an inclusive future. There are multiple steps to achieving diversity in dermatology. As mentioned in this article, some of these steps include focusing on specific changes within inclusivity in dermatology training, redesigning current diagnostic tools to decrease bias, and emphasizing the importance of cultural competency for dermatologists. To improve inclusivity in dermatology programs, specific schools have taken steps, including the following: diversity in clinical images and case studies, dermatoscopy and diagnostic techniques, cultural competence training, diverse faculty and leadership, and community outreach and engagement. These increase diversity and equity starting at the training level, improving and fostering trust between the healthcare workforce and the people right from the beginning. Without these steps though, the confidence that patients have in their doctors can significantly decrease, with the feeling that doctors don't truly understand their skin or hair type.

With most schools still not implementing measures to improve DEI, this delicate relationship is at risk. Current widely used diagnostic tools must also undergo reform. The current diagnostic accuracy of cutaneous/subcutaneous pathology in dark skin is, on average, only 44.3%. This is due to the major limitations in some of the most used classification systems, for instance, the FST. Understanding the limitations of these commonly used scales is essential to creating more effective systems that promote diversity and equitable care. For example, utilizing scales such as the DDI, the Robert Skin Classification System or even creating newer scales that address limitations in other scales, are all steps in the right direction to creating a reliable scale for all skin types. Both the DDI, and the Robert Skin Classification System emphasize their

focus on addressing the dearth of knowledge on skin of color, a limitation in the FST, and how both scales aim to meet the need for diversity as the diversity in racial and ethnic background grows in America.

Cultural competency is essential to achieving diversity in dermatology. By creating a workforce that is reflective of the population it serves, cultural and ethnic knowledge can be spread to reach and serve the needs of all patients. Furthermore, cultural competence in dermatology requires introspection on the current racial makeup of the dermatology field. Additionally, applying the various global approaches to dermatology in the United States can be pivotal. Teaching cultural competency can also prove crucial towards increasing inclusivity within dermatology. Focusing on specific key factors including diversifying the workforce, hosting panels, and increased exposure of skin of color images and case studies during training are small changes with a large impact. The importance of creating a diverse, inclusive community for dermatology to thrive can not be understated. Having equitable care should be a norm and patients should have confidence in their healthcare workforce, and the healthcare workforce should be confident in dealing with any patient, regardless of skin tone and racial/ethnic background. With collective and consistent efforts, the field of dermatology can become a diverse and equitable environment, where all patients are represented, and care is equitable.

Work Cited

Bellicoso, Emily, et al. "Diversity in Dermatology? An Assessment of Undergraduate Medical Education." *Journal of Cutaneous Medicine and Surgery*, vol. 25, no. 4, 2021, pp. 409–17. *PubMed*, https://doi.org/10.1177/12034754211007430.

Blewett, Lynn A., et al. "Patient Perspectives on the Cultural Competence of US Health Care Professionals." *JAMA Network Open*, vol. 2, no. 11, Nov. 2019, p. e1916105. *PubMed Central*, https://doi.org/10.1001/jamanetworkopen.2019.16105.

Bohler, Forrest, et al. "Analyzing Diversity, Equity, and Inclusion Content on Dermatology Fellowship Program Websites." *Medical Education Online*, vol. 29, no. 1, p. 2347762. *PubMed Central*, https://doi.org/10.1080/10872981.2024.2347762.

Cultural Competence: An Important Skill Set for the 21st Century. https://extensionpubs.unl.edu/publication/g1375/na/html/view. Accessed 3 Aug. 2024.

"Curriculum." *Penn Dermatology Residents*, https://dermatology.upenn.edu/residents/education/curriculum/. Accessed 9 July 2024.

Definition of Psoralen and Ultraviolet A Therapy - NCI Dictionary of Cancer Terms - NCI. 2 Feb. 2011, https://www.cancer.gov/publications/dictionaries/cancer-terms/def/psoralen-and-ultraviolet-a-therapy. nciglobal,ncienterprise.

Dermatologists: What Do They Do, Qualifications, and Procedures. https://www.medicalnewstoday.com/articles/286743. Accessed 5 Aug. 2024.

Disparities in Dermatology AI Performance on a Diverse, Curated Clinical Image Set | Science Advances. https://www.science.org/doi/10.1126/sciadv.abq6147. Accessed 5 Aug. 2024.

Embracing Diversity: New Skin Type System Addresses Procedures for Treating Skin of Color. https://www.dermatologytimes.com/view/embracing-diversity-new-skin-type-system-addresses-procedures-treating-skin-color. Accessed 5 Aug. 2024.

Fanous, Nabil. "A New Patient Classification for Laser Resurfacing and Peels: Predicting Responses, Risks, and Results." *Aesthetic Plastic Surgery*, vol. 26, no. 2, 2002, pp. 99–104. *PubMed*, https://doi.org/10.1007/s00266-002-1483-2.

Fernandez, Jesse. "Is It Hard to Become a Dermatologist? Obstacles and Rewards." *American University of Antigua*, 5 Dec. 2023, https://www.auamed.org/blog/is-it-hard-to-become-a-dermatologist/.

"Glogau Wrinkle Scale." *Dr. Richard Glogau*, https://sfderm.com/glogau-wrinkle-scale/. Accessed 22 Aug. 2024.

Goldman, Mitchell P. "Universal Classification of Skin Type." *Simplified Facial Rejuvenation*, edited by Melvin A. Shiffman et al., Springer, 2008, pp. 47–50. *Springer Link*, https://doi.org/10.1007/978-3-540-71097-4_3.

Inside the DEI Battle in Dermatology | Allure. https://www.allure.com/story/american-academy-of-dermatology-dei-resolution. Accessed 14 July 2024.

Kaundinya, Trisha, and Roopal V. Kundu. "Reply to Letter to the Editor 'Modifying a

Comprehensive Preclinical Dermatology Curriculum to Better Teach Dermatology in Skin of All Colors-Response to "Improving Cultural Competency and Practicing Cultural Humility in Dermatologic Training: Skin of Color Education and Board Certification."""" *Journal of the American Academy of Dermatology*, vol. 85, no. 6, Dec. 2021, p. e407. *PubMed*, https://doi.org/10.1016/j.jaad.2021.07.072.

Kawada, A. "UVB-Induced Erythema, Delayed Tanning, and UVA-Induced Immediate Tanning in Japanese Skin." *Photo-Dermatology*, vol. 3, no. 6, Dec. 1986, pp. 327–33.

Khalil, Nada, et al. "Undergraduate Dermatology Education: The Importance of Curriculum Review." *Skin Health and Disease*, vol. 3, no. 6, Sept. 2023, p. e288. *PubMed Central*, https://doi.org/10.1002/ski2.288.

Modified Fitzpatrick Scale-Skin Color and Reactivity - JDDonline - Journal of Drugs in Dermatology. https://jddonline.com/articles/modified-fitzpatrick-scale-skin-color-and-reactivity-S1545961623P0641X/. Accessed 5 Aug. 2024.

Nair, Lakshmi, and Oluwaseun A. Adetayo. "Cultural Competence and Ethnic Diversity in Healthcare." *Plastic and Reconstructive Surgery Global Open*, vol. 7, no. 5, May 2019, p. e2219. *PubMed Central*, https://doi.org/10.1097/GOX.0000000000002219.

Narla, Shanthi, et al. "Racial Disparities in Dermatology." *Archives of Dermatological Research*, vol. 315, no. 5, 2023, pp. 1215–23. *PubMed Central*, https://doi.org/10.1007/s00403-022-02507-z.

Okoji, U. K., et al. "Equity in Skin Typing: Why It Is Time to Replace the Fitzpatrick Scale." *The British Journal of Dermatology*, vol. 185, no. 1, July 2021, pp. 198–99. *PubMed*, https://doi.org/10.1111/bjd.19932.

Perlman, Katherine L., et al. "Skin of Color Lacks Representation in Medical Student Resources: A Cross-Sectional Study." *International Journal of Women's Dermatology*, vol. 7, no. 2, Jan. 2021, pp. 195–96. *PubMed Central*, https://doi.org/10.1016/j.ijwd.2020.12.018.

"Perspective: Racism in Academic Medicine Is Hindering Progress Toward Health Equity." *California Health Care Foundation*, https://www.chcf.org/publication/perspective-racism-academic-medicine-hindering-progress-toward-health-equity/. Accessed 3 Aug. 2024.

Roberts, Wendy E. "Skin Type Classification Systems Old and New." *Dermatologic Clinics*, vol. 27, no. 4, Oct. 2009, pp. 529–33. *ScienceDirect*, https://doi.org/10.1016/j.det.2009.08.006.

Sharma, Ajay N., and Bhupendra C. Patel. "Laser Fitzpatrick Skin Type Recommendations." *StatPearls*, StatPearls Publishing, 2024. *PubMed*, http://www.ncbi.nlm.nih.gov/books/NBK557626/.

Sonthalia, Sidharth, et al. "Dermoscopy Overview and Extradiagnostic Applications." *StatPearls*, StatPearls Publishing, 2024. *PubMed*, http://www.ncbi.nlm.nih.gov/books/NBK537131/.

U.S. Census Bureau QuickFacts: United States. https://www.census.gov/quickfacts/fact/table/US/PST045222. Accessed 5 Aug. 2024.

Ware, Olivia R., et al. "Racial Limitations of Fitzpatrick Skin Type." *Cutis*, vol. 105, no. 2, Feb. 2020, pp. 77–80.

Williams, Kiyanna, and Kanade Shinkai. "The Leaky Pipeline: A Narrative Review of Diversity in Dermatology." *Cutis*, vol. 109, no. 1, Jan. 2022, pp. 27–31. *PubMed*, https://doi.org/10.12788/cutis.0427.

Willis, I., and R. M. Earles. "A New Skin Classification System Relevant to People of African Descent." *Cosmetic Dermatology*, vol. 18, Mar. 2005, pp. 209–16.

Environmental and Neurobiological Factors Behind CTE Progression By Aditya Harathi

Abstract

The development of CTE may be impacted by environmental and neurobiological factors, which can influence treatment options for individuals with suspected CTE. Understanding factors that can affect neurodegeneration allows patients to make lifestyle modifications such as changing diet, maintaining consistent post-TBI checkups, and the location of which they live regarding air quality. Researching associated genes and proteins influencing neurodegeneration may help determine treatment development. Currently, there is no cure for CTE, and it can only be diagnosed via brain biopsy postmortem. However, early detection of typical symptoms like social withdrawal, depression, and aggression when coupled with the medical history of a patient with repetitive brain injuries can allow a patient to improve their quality of life by making modifications. Widespread awareness can encourage future studies on potential treatments targeting genes and innovative MRI imaging techniques to diagnose CTE *in vivo*.

Introduction

Chronic traumatic encephalopathy (CTE) is a neurodegenerative disease that results from repetitive traumatic brain injuries (TBIs) of varying severity. CTE alters a person's emotions, personality, and cognition. Four stages classify the severity of CTE in an individual. In Stage I, patients are typically asymptomatic, have slight memory deficits, and may develop feelings of depression (43). Stage II is characterized by further memory dysfunction and more charged emotions (43). Stage III patients exhibit more cognitive decline with more memory loss, decreased spatial awareness, and apathy (43). Stage IV has advanced cognitive decline similar to Parkinson's disease or Alzheimer's and patients present with more psychological problems (43). CTE was initially examined in sports such as boxing or football where repetitive hits occur often, however, CTE also has been noted to affect those in the armed forces due to concussive blasts from improvised explosive devices (IEDs) (45). CTE is characterized by the hyperphosphorylation of tau and the formation of neurofibrillary tangles (NFTs) which disrupt neuronal communication leading to neurodegeneration (43,44). It is not clear why some individuals affected by CTE are less affected than others. By better understanding the factors that could influence the progression of CTE within an individual, there would be a better chance of developing a way to slow its development. Recent studies have investigated proteins such as Pin1 and methods to diagnose the disease *in vivo* via MRI (46,50). Environmental factors that could impact development include but are not limited to diet, access to healthcare resources, and environmental pollutants. Neurobiological factors such as the *ApoE2/Apoe4* gene expression, tau phosphorylation, and microglia may also affect the progression of CTE. By understanding factors such as these, people who are at risk for developing CTE or suspect its development may improve their prognosis and quality of life.

Environmental Factors

The progression of CTE and neurodegenerative diseases alike can be affected by the environment of a patient. Environmental factors can include diet, access to healthcare, and pollution.

Diet

When considering diet, one study found that dementia patients who consumed more than 7 grams (g) of olive oil per day for 28 years had a 28% decreased mortality rate when compared to dementia patients who consumed less than 7g of olive oil per day (1). This suggests that there is a modifiable environmental factor to reduce mortality in dementia patients. Additionally, to research the connection between ultra-processed foods (UPFs) and cognitive decline (6,7), found that those whose daily calories composed of 19.9% UPFs or greater experienced a greater rate of cognitive decline than those whose diets comprised less than 19.9% UPFs over the same time (6). Such cognitive decline is correlated with increased inflammation within the brain (6). This inflammation, as another study suggested, could be the result of high percentages of UPFs within one's diet (27). The amount of UPFs in a diet can be mitigated by developing healthy eating habits by replacing UPFs with healthier options such as fish (6) which contains omega-3 fatty acids, which decrease the rate at which neurons die and lower neuroinflammation (8), both of which are accelerated in dementia patients (12). Neuroinflammation promotes the acceleration of tau accumulation (2,12) which is the cornerstone of CTE pathology; which is marked by perivascular tau accumulation in the sulci of the brain (2,13). Perhaps, modifying diet by reducing UPFs may decrease the risk of tau accumulation or slow the progression of CTE. Another study highlighted that a high sodium diet (HSD) correlated to a decrease in brain blood flow within subject mice compared to mice fed standard rodent chow (9). This reduction in blood flow led to deficits in object recognition similar to the deficits found in dementia. Due to the cerebral hypoperfusion, endothelial dysfunction occurred as endothelial nitric oxide (eNOS) production was disrupted (18). Normally, eNOS is responsible for maintaining cardiovascular homeostasis and some neurovascular protections (10,17,18,28), however due to its absence tau hyperphosphorylation can and did occur in the case of the mice. Therefore, when considering CTE patients, an HSD could worsen symptoms by accelerating neurodegeneration and neurofibrillary tangle (NFT) production via the impact on eNOS as shown with the mice. There are more parts to dieting and it varies from person to person. When taking in all aspects of diet, it can potentially impact the progression of diseases such as CTE, therefore by being mindful of what one is eating, they may have a more favorable prognosis.

Access to Healthcare

CTE is attributed to repetitive trauma to the brain, therefore the level of care and access to such care that patients receive following TBIs could play a role in modifying risk factors (2). In a study regarding the follow-up care of 831 mTBI patients transported to a level 1 trauma center immediately after their injury and were stabilized; 48% of patients who presented with severe postconcussive symptoms were not met by a physician after 3 months of their injury's

occurrence (3). Those who did visit the trauma center regularly within 3 months of their injury had better recovery outcomes compared to their counterparts who did not follow up regularly. Perhaps, by encouraging patients to make multiple visits in the months following their injury, physicians can treat them more effectively and educate them on good practices to avoid having future concussions from occurring and prevent the worsening of their symptoms as they recover. The recovery process following a TBI can involve many different departments (42) and as such each patient's rehab must be treated differently, but monitored closely nonetheless. Furthermore, a physician's vigilance and a patient's cooperation can decrease the patient's chances of developing CTE. Additionally, prehospital care plays a crucial role in a patient's condition, specifically in the time it takes for a TBI patient to be transported to an appropriate medical facility. Approximately half the deaths caused by TBIs occur within the first 2 hours of injury (11), as such, rapid transport and appropriate treatment, such as C-Spine immobilization, by EMS personnel immediately following the injury is crucial for preventing further deterioration of the patient's health. Many mTBI and severe TBI patients will be discharged to outpatient centers to receive care and physical therapy after their initial visit to the hospital. There they may need rehabilitation both mentally and physically. Still, sometimes these outpatient centers will not have the same standards for treating TBI patients or the proper resources to do so such as enough beds, rooms, and mental health services (16). The reason the accessibility and level of care given to patients following a TBI injury is so important is to prevent further damage to the spinal cord, the brain, and ultimately the impact it has on the patient's quality of life. Without a proper medical evaluation, for example, a football player who may have taken a hard hit to the head in a game may be put back in. This results in increased chances of reinjury especially since TBI patients present symptoms including dizziness, blurred vision, and confusion(15) and will most likely not be as aware of the potential threats surrounding them. Also, in cases of domestic violence or intimate partner violence (IPV) which affects millions of people around the world, a majority of whom are women, many situations are not reported due to the stigma surrounding these traumas(19). By destigmatizing IPV and bringing awareness to the brain damage caused by repetitive TBIs sustained, more victims will receive proper care for their injuries and thus prevent the worsening of their symptoms. Repetitive TBIs are a major factor in developing CTE(5) and receiving the necessary care following the injury in situations like these is why access to healthcare resources is important.

Pollution

In addition to diet and healthcare, ambient pollution may be a notable factor in CTE progression. Particulate matter with a diameter of less than 2.5 μm (PM2.5) are ultrafine pollutants that can cross the blood-brain barrier (BBB) in the brain due to their microscopic size (49). The BBB is composed primarily of astrocytes that line the brain's endothelium and is a semipermeable membrane that allows certain molecules and substances to cross into the brain (14). However, when unwanted particles such as PM2.5 pollutants enter the BBB, they can cause oxidative stress (4,14,23). Oxidative stress is the imbalance of free radicals and antioxidants

within the body and is important when looking at neurodegenerative disease progression. Free radicals can be formed from metabolic pathways or outside sources - i.e. pollutants such as PM2.5 (48) - and oxidative stress causes microglial activation and further activation of pro-inflammatory cytokines (24). This process, to rid unnecessary particles, accelerates brain inflammation, and microglial activation in this manner is a hallmark of AD disease progression (24). This is similar to another neurodegenerative disease, CTE, where the hyperphosphorylation of tau leads to microglial activation to clean up said tau, however, results in brain inflammation (31). Compared to diet and access to healthcare following TBIs, pollution is the harder environmental factor of CTE discussed to change in one's life due to more outside input than inside. However, small changes whether its habits of smoking, leaving cars running in closed spaces, or big changes such as moving to areas with higher air quality will benefit rather than harm.

Accounting for the environmental factors that can affect CTE progression, further strides in CTE prevention can be made by society recognizing that the prognosis of this disease and potential diagnosis can be improved by making mindful decisions regarding lifestyle decisions and obtaining the proper care after TBIs.

Neurobiological Factors

Neurobiological factors may also affect CTE progression in individuals. These can include genetic factors, such as the variants of the Apolipoprotein E alleles carried within an individual, or nongenetic factors, such as tau isoforms and microglia.

ApoE2/ApoE4

Apolipoprotein E (ApoE) is a protein responsible for lipid metabolism functions in the brain and has three forms: ApoE2, ApoE3, and ApoE4 (20). In the literature, ApoE4 has been associated with an increased chance of developing neurodegenerative diseases such as AD (20). Whereas, the presence of ApoE2 or ApoE3 is cited with a decreased chance of developing AD when compared to ApoE4 (21). To show the relevance of ApoE and the progression of neurodegenerative diseases, a post-mortem analysis of mouse brains revealed that carriers of one *ApoE4* allele had a 14% increased rate of neurodegeneration while two *ApoE4* alleles increased the rate of progression by 23% when compared to non-carriers (22), suggesting that the presence of *ApoE4* increases the risk of neurodegenerative disease in a dose-dependent manner. Expression of *ApoE4* in older mice demonstrated significantly more BBB breakdown when compared to *ApoE3*-expressing mice, regardless of any protein pathologies (32). BBB breakdown can lead to cognitive decline and tauopathy (32,33). Therefore, *ApoE4* may contribute to the progression of tauopathies such as CTE.

Tau Phosphorylation

Tau is a protein that stabilizes microtubules and regulates axonal transport (25). Six tau isoforms in the adult brain (25) are generated by exon excision during splicing (26). Each tau

isoform contains an N terminus (N), proline-rich domain, microtubule-binding domain, and C terminus (25). The inclusion of exons 2 and 3 allows for 0N, 1N, or 2N, respectively, to be included within the tau isoform. When exon 10 is included, 3 repeats (3R) occur within the microtubule-binding, however, when exon 10 is excluded, 4 repeats occur (4R) (25). When looking at tau pathology in deceased patients diagnosed in life with AD and diagnosed post-mortem with CTE, a group of researchers found 3R and 4R isoforms of tau present in varying combined amounts (36). They also noted, that as the pathological intensity decreased, the 4R isoform ratio to the 3R increased. However, when the tau burden increased, the 3R isoform ratio to the 4R increased. Therefore, in the early stages of CTE, 4R tau is more significant in disease progression, while in the end stages, 3R tau takes more part in the processes. Tau may not be the initial cause behind neurodegeneration in diseases such as CTE and AD, however, when it gets altered and induces tauopathy, it does irreparable damage (25,36). In CTE, tau accumulates perivascularly in the sulci regions of the brain (2,5,13,15,25,34,36). NFTs form as tau aggregation occurs, and nerve cells are obstructed from communicating with each other due to their placement (2,5,13,35,39).

Microglia

Microglia are cells that function as the immune cells of the nervous system. They phagocytose debris and dead cells within the nervous tissue (29). During TBIs, microglia are activated as part of an immune response within the brain to clear debris in the affected areas (30). Supposedly, this microglial activation results in acute inflammation of the injured area until microglia are not needed. However, one study found that this microglial activation lasted for one year in mice tested using controlled cortical impacts (CCI) and in the same span the researchers found that neuroinflammation was still occurring (31). This is due to microglial producing pro-inflammatory responses in humans and mice alike, which can last for some time after mTBIs and severe TBIs (31). The initial inflammation may be beneficial, however, when prolonged it becomes detrimental to the patient's brain due to the increasing neurotoxicity of these pro-inflammatory responses over time. Therefore, long periods of microglial activation could be devastating to patients recovering from a TBI and those who experience repetitive TBIs over a lifetime as is often with CTE.

Taking these neurobiological factors together can help give way to a further understanding of CTE disease pathology and lead to innovation in the research surrounding it as well as progress toward the care of the patients with suspected CTE.

Potential Treatment and Imaging Methods

Pin1

Currently, there is no known cure for CTE. However, several studies have been investigating proteins that may play a role in the pathogenesis of neurodegenerative disease. A

deficiency or depletion of Pin1 has been associated with neurodegenerative diseases such as AD and FTD (38).

Generally, Pin1 plays a role in cell division, and apoptosis, and regulates phosphorylation of multiple proteins (38). Pin1 isomerizes tau from the *cis* form to the *trans* form at serine/threonine phosphorylation sites at least two sites (37,40). Pin1 has been shown to promote phosphorylated tau (ptau) dephosphorylation and promote tau's normal function of microtubule assembly and support (37). Pin1's ability to inhibit the degeneration of the brain due to tau phosphorylation in tauopathies could lead to treatments targeting Pin1 in CTE. Another study explored the protein tacrolimus (FK506) in mice whom it was put in via food pellets, which can provoke the Pin-1 signaling cascade and has been linked to reduced AD prevalence in organ transplant recipients who this protein was initially approved for use (41). The introduction of FK506 in mice in the study led to a 35% increase in Pin1 activity within them. This activity was paired with no negative effects such as tumor growth which could be a result of elevated Pin1 levels (38). Given that Pin1 is shown in lower amounts in individuals with neurodegenerative diseases such as AD (38) and that the use of FK506 is attributed to decreased AD pathology in individuals in who it was used (41) while it also increases Pin1 activity, it could be assumed that a method of treatment utilizing the two proteins could be used in CTE treatment due to Pin1's role in ptau dephosphorylation (37-40).

MRI

While magnetic resonance imaging (MRI) can not be used to diagnose CTE in a deceased patient, it can be used to view the atrophy of an autopsy-confirmed CTE patient (46). One study viewed the correlation between ptau and atrophy in the brains of such patients. The study viewed the brains of 31 donors who had normal cognition at the time of death and 55 donors without normal cognition at the time of death with 44/55 having Stage III/Stage IV CTE. The study found a prominent accumulation of tau in the frontal and temporal lobe which has been previously associated with CTE-diagnosed patients (2,5,13,46). Also, among the donors with CTE, the level of atrophy correlated to the severity of ptau accumulation (46). This finding is consistent with previous discussions of CTE, therefore, there may be a potential to utilize MRI imaging to understand the disease pathology *in vivo* in suspected patients to diagnose *in vivo*. This is speculation for now, however, with Boston University citing a similar hope and using this study to back its claim there is a real potential for further studies using MRI to examine CTE paired with innovation to reach a point where CTE can be diagnosed *in vivo* (50).

Discussion

This paper explores the various factors influencing the progression of CTE and examines potential strategies for mitigating their impact to improve prognosis. Additionally, while acknowledging the lack of existing treatments, a deeper understanding of the genetic factors underlying CTE could pave the way for both generalized therapies and individualized treatment approaches. Since CTE can only be definitively diagnosed postmortem through brain biopsy,

advancements in the study of neurodegenerative diseases and tauopathies may facilitate earlier detection. This could be achieved either by improving current medical imaging techniques (46) or by developing novel biomarkers capable of tracking tauopathy in individuals with neurodegenerative conditions, including but not limited to CTE. Although CTE is commonly associated with individuals in combat sports or high-impact careers, it is crucial to recognize the role of subconcussive impacts in disease development across various populations. Subconcussive hits, or mild traumatic brain injuries (mTBIs), are approximately 1.5 times less damaging than concussive blows but occur 400 times more frequently (47). The high frequency of these impacts contributes to the excessive neuroinflammation observed in CTE-affected brains and plays a key role in the hyperphosphorylation of tau, a hallmark of disease progression. Creating awareness around the circumstances by which CTE may develop can improve the odds of various populations who are clearly or not so clearly deemed susceptible to it. This can be done by better understanding how an individual's lifestyle or genetics can impact their prognosis to innovate current technologies and methods in their favor.

Conclusion

CTE has various factors that may affect its progression within an individual. By better understanding these factors, we may develop methods to manage its progression and improve the prognosis of patients suspected of having CTE and related diseases.

Works Cited

Tessier A, Cortese M, Yuan C, et al. Consumption of Olive Oil and Diet Quality and Risk of Dementia-Related Death. *JAMA Netw Open.* 2024;7(5):e2410021. doi:10.1001/jamanetworkopen.2024.10021

Moore, Kendall B E et al. "Hyperphosphorylated tau (p-tau) and drug discovery in the context of Alzheimer's disease and related tauopathies." *Drug discovery today* vol. 28,3 (2023): 103487. doi:10.1016/j.drudis.2023.103487

Seabury SA, Gaudette É, Goldman DP, et al. Assessment of Follow-up Care After Emergency Department Presentation for Mild Traumatic Brain Injury and Concussion: Results From the TRACK-TBI Study. *JAMA Netw Open.* 2018;1(1):e180210. doi:10.1001/jamanetworkopen.2018.0210

Vanbrabant K, Van Dam D, Bongaerts E, et al. Accumulation of Ambient Black Carbon Particles Within Key Memory-Related Brain Regions. *JAMA Netw Open.* 2024;7(4):e245678. doi:10.1001/jamanetworkopen.2024.5678

McKee, Ann C et al. "The neuropathology of chronic traumatic encephalopathy." *Brain pathology (Zurich, Switzerland)* vol. 25,3 (2015): 350-64. doi:10.1111/bpa.12248

Gomes Gonçalves N, Vidal Ferreira N, Khandpur N, et al. Association Between Consumption of Ultraprocessed Foods and Cognitive Decline. *JAMA Neurol.* 2023;80(2):142–150. doi:10.1001/jamaneurol.2022.4397

Li, Jian-Guo et al. "The Anti-Neuroinflammatory Effect of Extra-Virgin Olive Oil in the Triple Transgenic Mouse Model of Alzheimer's Disease." *Journal of Alzheimer's disease : JAD* vol. 100,1 (2024): 119-126. doi:10.3233/JAD-240374

Dighriri, Ibrahim M et al. "Effects of Omega-3 Polyunsaturated Fatty Acids on Brain Functions: A Systematic Review." *Cureus* vol. 14,10 e30091. 9 Oct. 2022, doi:10.7759/cureus.30091

Faraco et al. Dietary salt promotes neurovascular and cognitive dysfunction through a gut-initiated TH17 response. Nature Neuroscience. January 15, 2018. doi: 10.1038/s41593-017-0059-z

Chen, Kejing et al. "Nitric oxide in the vasculature: where does it come from and where does it go? A quantitative perspective." *Antioxidants & redox signaling* vol. 10,7 (2008): 1185-98. doi:10.1089/ars.2007.1959

National Academies of Sciences, Engineering, and Medicine, Health and Medicine Division, Board on Health Care Services, Board on Health Sciences Policy, Committee on Accelerating Progress in Traumatic Brain Injury Research and Care, editors. *Traumatic Brain Injury: A Roadmap for Accelerating Progress*. National Academies Press, 2022. Chapter 5, "Acute Care After Traumatic Brain Injury," https://www.ncbi.nlm.nih.gov/books/NBK580074/.

Gaikwad, Sagar et al. "Senescence, brain inflammation, and oligomeric tau drive cognitive decline in Alzheimer's disease: Evidence from clinical and preclinical studies."

Alzheimer's & dementia : the journal of the Alzheimer's Association vol. 20,1 (2024): 709-727. doi:10.1002/alz.13490

McKee, Ann C et al. “Chronic traumatic encephalopathy (CTE): criteria for neuropathological diagnosis and relationship to repetitive head impacts.” *Acta neuropathologica* vol. 145,4 (2023): 371-394. doi:10.1007/s00401-023-02540-w

Wang, Jian et al. “The Impact of Air Pollution on Neurodegenerative Diseases.” *Therapeutic drug monitoring* vol. 43,1 (2021): 69-78. doi:10.1097/FTD.0000000000000818

Cleveland Clinic, “Traumatic Brain Injury.” *Cleveland Clinic,* 25 Jan. 2025, https://my.clevelandclinic.org/health/diseases/8874-traumatic-brain-injury#outlook-prognosis

"6 Rehabilitation and Long-Term Care Needs After Traumatic Brain Injury." National Academies of Sciences, Engineering, and Medicine. 2022. Traumatic Brain Injury: A Roadmap for Accelerating Progress. Washington, DC: The National Academies Press. doi: 10.17226/25394.

Tewari, Devesh et al. “Role of Nitric Oxide in Neurodegeneration: Function, Regulation, and Inhibition.” *Current neuropharmacology* vol. 19,2 (2021): 114-126. doi:10.2174/1570159X18666200429001549

Faraco, G., Hochrainer, K., Segarra, S.G. *et al.* Dietary salt promotes cognitive impairment through tau phosphorylation. *Nature* 574, 686–690 (2019). https://doi.org/10.1038/s41586-019-1688-z

Valera, Eve. “Intimate partner violence and traumatic brain injury: An invisible public health epidemic.” *Harvard Medical School, Harvard Health Publishing, 17 March 2022,* https://www.health.harvard.edu/blog/intimate-partner-violence-and-traumatic-brain-injury-an-invisible-public-health-epidemic-201812132708

Huang, Yadong, and Robert W Mahley. “Apolipoprotein E: structure and function in lipid metabolism, neurobiology, and Alzheimer's diseases.” *Neurobiology of disease* vol. 72 Pt A (2014): 3-12. doi:10.1016/j.nbd.2014.08.025

Serrano-Pozo, Alberto et al. “APOE and Alzheimer's disease: advances in genetics, pathophysiology, and therapeutic approaches.” *The Lancet. Neurology* vol. 20,1 (2021): 68-80. doi:10.1016/S1474-4422(20)30412-9

Shi, Yang et al. “ApoE4 markedly exacerbates tau-mediated neurodegeneration in a mouse model of tauopathy.” *Nature* vol. 549,7673 (2017): 523-527. doi:10.1038/nature24016

de Prado Bert, Paula et al. “The Effects of Air Pollution on the Brain: a Review of Studies Interfacing Environmental Epidemiology and Neuroimaging.” *Current environmental health reports* vol. 5,3 (2018): 351-364. doi:10.1007/s40572-018-0209-9

Onyango, Isaac G et al. “Neuroinflammation in Alzheimer's Disease.” *Biomedicines* vol. 9,5 524. 7 May. 2021, doi:10.3390/biomedicines9050524

Noble, Wendy et al. “The importance of tau phosphorylation for neurodegenerative diseases.” *Frontiers in neurology* vol. 4 83. 1 Jul. 2013, doi:10.3389/fneur.2013.00083

"Exon." *National Human Genome Research Institute, 1 Jan. 2025, https://www.genome.gov/genetics-glossary/Exon*

Mannino, A., Daly, A., Dunlop, E. *et al.* Higher consumption of ultra-processed foods and increased likelihood of central nervous system demyelination in a case-control study of Australian adults. *Eur J Clin Nutr* 77, 611–614 (2023). https://doi.org/10.1038/s41430-023-01271-1

Katusic, Zvonimir S et al. "Emerging Roles of Endothelial Nitric Oxide in Preservation of Cognitive Health." *Stroke* vol. 54,3 (2023): 686-696. doi:10.1161/STROKEAHA.122.041444

Rogers, Kara. "microglia". Encyclopedia Britannica, 11 Dec. 2024, https://www.britannica.com/science/microglia. Accessed 1 January 2025.

Makinde, Hadijat M et al. "The Role of Microglia in the Etiology and Evolution of Chronic Traumatic Encephalopathy." *Shock (Augusta, Ga.)* vol. 48,3 (2017): 276-283. doi:10.1097/SHK.0000000000000859

Loane, David J et al. "Progressive neurodegeneration after experimental brain trauma: association with chronic microglial activation." *Journal of neuropathology and experimental neurology* vol. 73,1 (2014): 14-29. doi:10.1097/NEN.0000000000000021

Montagne, A., Nikolakopoulou, A.M., Huuskonen, M.T. *et al. APOE4* accelerates advanced-stage vascular and neurodegenerative disorder in old Alzheimer's mice via cyclophilin A independently of amyloid-β. *Nat Aging* 1, 506–520 (2021). https://doi.org/10.1038/s43587-021-00073-z

Montagne, Axel et al. "APOE4 leads to blood-brain barrier dysfunction predicting cognitive decline." *Nature* vol. 581,7806 (2020): 71-76. doi:10.1038/s41586-020-2247-3

Wang, Y., Mandelkow, E. Tau in physiology and pathology. *Nat Rev Neurosci* 17, 22–35 (2016). https://doi.org/10.1038/nrn.2015.1

Chen, Liam. "What triggers tauopathy in chronic traumatic encephalopathy?." *Neural regeneration research* vol. 13,6 (2018): 985-986. doi:10.4103/1673-5374.233439

Cherry, Jonathan D et al. "Tau isoforms are differentially expressed across the hippocampus in chronic traumatic encephalopathy and Alzheimer's disease." *Acta neuropathologica communications* vol. 9,1 86. 12 May. 2021, doi:10.1186/s40478-021-01189-4

Lu KP, Kondo A, Albayram O, Herbert MK, Liu H, Zhou XZ. Potential of the Antibody Against *cis*–Phosphorylated Tau in the Early Diagnosis, Treatment, and Prevention of Alzheimer Disease and Brain Injury. *JAMA Neurol.* 2016;73(11):1356–1362. doi:10.1001/jamaneurol.2016.2027

Chen, Y., Wu, Yr., Yang, Hy. *et al.* Prolyl isomerase Pin1: a promoter of cancer and a target for therapy. *Cell Death Dis* 9, 883 (2018). https://doi.org/10.1038/s41419-018-0844-y

Zhang, Xue et al. "Tau Pathology in Parkinson's Disease." *Frontiers in neurology* vol. 9 809. 2 Oct. 2018, doi:10.3389/fneur.2018.00809

Kimura, Taeko et al. "Isomerase Pin1 stimulates dephosphorylation of tau protein at cyclin-dependent kinase (Cdk5)-dependent Alzheimer phosphorylation sites." *The*

Journal of biological chemistry vol. 288,11 (2013): 7968-7977. doi:10.1074/jbc.M112.433326

Stallings, N.R., O'Neal, M.A., Hu, J. *et al.* Long-term normalization of calcineurin activity in model mice rescues Pin1 and attenuates Alzheimer's phenotypes without blocking peripheral T cell IL-2 response. *Alz Res Therapy* 15, 179 (2023). https://doi.org/10.1186/s13195-023-01323-5

"Rehabilitation After Traumatic Brain Injury." *Johns Hopkins Medicine,* Accessed: 1 Jan. 2025, https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/rehabilitation-after-traumatic-brain-injury

Fesharaki-Zadeh, Arman. "Chronic Traumatic Encephalopathy: A Brief Overview." *Frontiers in neurology* vol. 10 713. 3 Jul. 2019, doi:10.3389/fneur.2019.00713

Atherton K, Han X, Chung J, et al. Association of *APOE* Genotypes and Chronic Traumatic Encephalopathy. *JAMA Neurol.* 2022;79(8):787–796. doi:10.1001/jamaneurol.2022.1634

Catanese, Lisa. "What is CTE? Understanding chronic traumatic encephalopathy." *Harvard Medical School, Harvard Health Publishing, 28 March 2024, https://www.health.harvard.edu/mind-and-mood/what-is-cte-understanding-chronic-traumatic-encephalopathy#:~:text=Chronic%20traumatic%20encephalopathy%20was%20first%20identified%20in%20boxers,this%20type%20of%20TBI%20are%20professional%20football%20players*

"MRIs May be Initial Window into CTE Diagnosis in Living; Approach May Shave Years Off Diagnosis." *Boston University Chobanian & Avedisian School of Medicine,* https://www.bumc.bu.edu/camed/2021/12/08/mris-may-be-initial-window-into-cte-diagnosis-in-living-approach-may-shave-years-off-diagnosis/

Nowinski, Christopher J et al. "'Subconcussive' is a dangerous misnomer: hits of greater magnitude than concussive impacts may not cause symptoms." *British journal of sports medicine* vol. 58,14 754-756. 1 Jul. 2024, doi:10.1136/bjsports-2023-107413

Ganguly, Upasana et al. "Oxidative Stress, Neuroinflammation, and NADPH Oxidase: Implications in the Pathogenesis and Treatment of Alzheimer's Disease." *Oxidative medicine and cellular longevity* vol. 2021 7086512. 16 Apr. 2021, doi:10.1155/2021/7086512

Block, Michelle L, and Lilian Calderón-Garcidueñas. "Air pollution: mechanisms of neuroinflammation and CNS disease." *Trends in neurosciences* vol. 32,9 (2009): 506-16. doi:10.1016/j.tins.2009.05.009

Thurston, Andrew, Gina, Degravio. "8 Major Findings and Headlines from BU CTE Researchers in the Past Year." *The Brink, Boston University, 24 May 2022,* https://www.bu.edu/articles/2022/8-major-findings-from-bu-cte-researchers-last-year/

From Impact to Recovery: Unraveling the Journey of Traumatic Brain Injuries
By Atharv Mane

Abstract

Traumatic brain injuries (TBIs) occur when an individual experiences external force to the head, causing damage with varying severity and affecting different brain regions. While no FDA-approved treatments currently exist, promising therapies like amantadine and memantine are used to treat injuries. This review synthesizes peer-reviewed research on TBI types, ranging from mild concussions to severe herniations, highlighting their diverse outcomes and exploring current management strategies. The goal of this work is to contribute to the advancement of new therapies.

Introduction

Traumatic brain injuries (TBIs) are a significant public health concern, arising from external forces that disrupt normal brain function (Capizzi, Woo, & Verduzco-Gutierrez, 2020). These injuries can result from various incidents, including vehicular accidents, falls, sports injuries, and assaults (Mckee & Daneshvar, 2015). TBIs disproportionately affect older adults (aged 75+) and young children (aged 0-4), who experience the highest rates of TBI-related hospital visits (Capizzi, Woo, & Verduzco-Gutierrez, 2020). There are two types of TBIs: closed TBIs (non-penetrative) (National Academies Press, 2009) and open TBIs (penetrating) (Alao, Munakomi, & Waseem, 2024). Closed TBIs are caused when the external trauma to the head is strong enough to shake the brain inside the skull (National Academies Press, 2009). These types of TBIs can be caused by vehicle accidents, explosions, and sports injuries (National Academies Press, 2009). Open TBIs are when projectiles or objects in motion acted upon by gravity, pierce the individual's head and directly damage the skull and brain tissue (Alao, Munakomi, and Waseem, 2024).

TBIs also differ in severity, ranging from mild to moderate or severe TBIs. There is a metric used to determine the severity of a TBI, called the Glasgow Coma Scale. On a scale of 3-15, mild TBIs scale from 14-15, moderate TBIs scale from 9-13, and severe TBIs scale from 3-8 (Mena et al., 2011). Mild TBIs (mTBI) are the most common type reported, with 90% being mild (Bielanin, Metwally, Paruchuri, & Sun, 2023). One subset of mTBIs is concussions (Maas, Menon, Manley, & Abrams, 2022). Concussions are ubiquitous in sports like boxing, American football, and various other contact sports (Ling, Hardy, & Zetterberg, 2015). Moderate and severe TBIs are commonly grouped and can cause permanent damage. This paper will seek to accomplish several goals, such as providing a detailed scope into the many types of traumatic brain injuries, compiling existing research about the effects of traumatic brain injuries on the head, discussing common treatment options, branching off into similar types of injuries and neurological disorders, and provide further understanding into the injured brain and contribute to improved therapies.

mTBIs and Associated Risks

Traumatic brain injuries range in severity from mild concussions to severe traumatic brain herniations, therefore, these injuries have varying adverse effects on the brain. Starting with mild traumatic brain injuries (mTBIs), these injuries stem from contact sports like boxing and American football (Ling, Hardy, & Zetterberg, 2015). They result when the individual receives a mild jolt or blow to the head, resulting in an injury to the brain. Some symptoms include headaches, nausea, fatigue, drowsiness, dizziness, or amnesia (Ling, Hardy, & Zetterberg, 2015). After the injury, the individual can experience cognitive, behavioral, or mental symptoms such as loss of consciousness for up to a few minutes, memory/concentration issues, mood swings/changes, feelings of depression and anxiety, sleep difficulties or even sleeping more than usual (Skjeldal, Skandsen, Kinge, Glott, & Solbakk, 2022). Another case of mTBIs is skull fractures (Cadman & Han, 2024), although they can vary from mild to life-threatening. Classified as a bone break, it is commonly caused during car accidents, sports injuries, heavy falls, and physical assault (Cadman & Han, 2024).

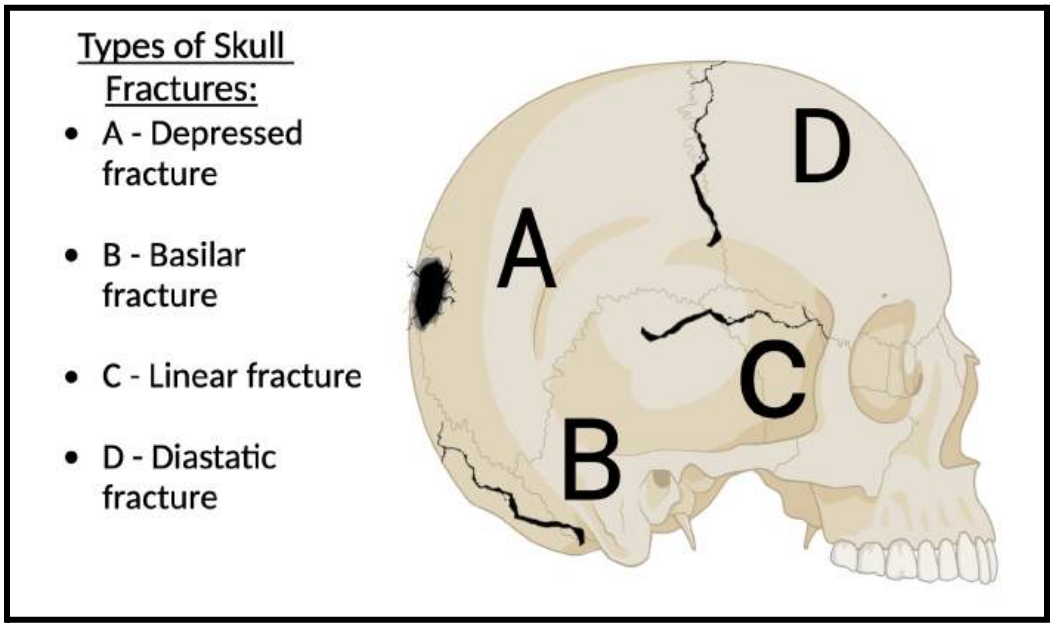


Figure 1: An illustration of different types of traumatic brain injury, highlighting the location and extent of damage. Created with BioRender.

There are different types of skull fractures: linear fractures (McGrath & Taylor, 2023), depressed fractures (McGrath & Taylor, 2023), basal fractures (McGrath & Taylor, 2023) and diastatic fractures (McGrath & Taylor, 2023). Linear fractures are the most common type of skull fractures, which usually happen in the temporal and parietal region, which is the region above your ears (McGrath & Taylor, 2023). Furthermore, a depressed fracture is a break in your skull that strikes your head directly and requires a neurosurgical opinion (McGrath & Taylor, 2023). The basal, or basilar fracture, is when the bone breaks at the base of your skull, which includes the bones behind the individual's face (McGrath & Taylor, 2023). This fracture is rather

complex, for the underlying cranial nerves and sinuses that might be affected. As a result, hearing loss, paralysis, and a decreased sense of smell. Finally, a diastatic fracture is a separation of the cranial sutures (fibrous joints connect the bones of the skull) (Bell, 2024). Although the signs of a skull fracture vary based on the type of skull fracture, a break might be evident if a lump or swelling is present on the head, cerebrospinal fluid leaking from the nose and ears, or blood running from your nose and ears, or from a wound near the injury site (Mao, 2023). Additionally, some symptoms include nausea, vomiting, confusion, vision loss, loss of feeling in various body parts, balance and coordination difficulty, fatigue, and slurred speech (Mao, 2023). Given prior knowledge, one can be equipped to recognize a major injury like a skull fracture, but if left unchecked, the fracture can create serious complications.

Intracranial Hemorrhages and Associated Risks

Although skull fractures can vary in severity, they can be catalysts to more serious TBIs, which makes early identification of a fracture paramount to an individual's future. As a prominent example, a skull fracture can advance into an intracranial hemorrhage, which is classified as bleeding in the brain tissue (Tenny & Thorell, 2024). The bleeding in the brain cuts off oxygen to the brain, further identifying an intracranial hemorrhage as a type of stroke. There are four types of intracranial hemorrhaging: epidural hemorrhage, subdural hemorrhage, subarachnoid hemorrhage, and intraparenchymal hemorrhage (Tenny & Thorell, 2024).

Firstly, epidural hemorrhaging (EH) ranges from moderate to severe and is classified into arterial epidural hemorrhaging and venous epidural hemorrhaging (Tenny & Thorell, 2024). Arterial epidural hemorrhaging (AEH) is caused by blunt head trauma in the temporal region of the head (Tenny & Thorell, 2024), damaging the meningeal artery, which results in arterial bleeding (Khairat & Waseem, 2023). As a result, bleeding occurs within the potential space between the inner table of the skull and the outer layer of the dura mater (Khairat & Waseem, 2023). Additionally, this type of injury accounts for 10% of TBIs requiring hospitalization, and 15% of all fatal head injuries (Khairat & Waseem, 2023). Moving on to venous epidural hemorrhaging (VEH) occurs when a dural venous sinus is lacerated, which then follows into venous bleeding. This injury is present beneath a fracture of the squamous part of the temporal bone of the skull (Khairat & Waseem, 2023). Although 75% of VEHs occur in the temporal region among adults, the occipital, frontal, and posterior fossa regions among children (Khairat & Waseem, 2023). Furthermore, this type of hemorrhaging makes up 10% of epidural hemorrhaging (Khairat & Waseem, 2023).

Secondly, subdural hemorrhaging (SH) is a severe TBI that occurs when bleeding occurs in the subdural space of the head (Sharma, 2024). Moreover, SH is believed to be caused by the stretching and tearing of bridging cortical veins as they cross into the subdural space to drain into an adjacent dural sinus (Sharma, 2024). The reason these veins tear is associated with a sudden change in the velocity of the head (Sharma, 2024). Additionally, the arachnoid may be damaged as well, which may lead to a mixture of cerebrospinal fluid and blood in the subdural space (Sharma, 2024). The arachnoid is located between the dura mater and the pia mater which is

involved in cerebrospinal fluid metabolism via the subarachnoid space (Ghannam & Kharazi, 2023).

Thirdly, a subarachnoid hemorrhage (SAH) is a relatively mild TBI and occurs when a medical aneurysmal rupture occurs because of a TBI, causing bleeding in the subarachnoid space between the arachnoid membrane and the pia mater (Kairys, Das, & Garg, 2022). SAHs can be either traumatic or nontraumatic (Kairys, Das, & Garg, 2022), but for the sake of this review, I will only review the traumatic aspect of this type of hemorrhaging. SAHs take place near the site of traumatic injury, with radiological clues including localized bleeding in a superficial sulcus (a depressed groove in the cerebral cortex), and adjacent skull fractures (Kairys, Das, & Garg, 2022). SAHs are devastating, for almost half of the individuals who are affected by an aneurysm caused by SAHs don't survive within thirty days, while only one-third of individuals who survive have complications (Kairys, Das, & Garg, 2022).

Fourthly, an intraparenchymal hemorrhage (IPH) is a severe TBI that occurs when bleeding takes place in the brain tissue itself (Pavel). Additionally, traumatic events that occur regarding IPH cause mechanical damage to the spinal parenchyma at the lesion center. The size of this hemorrhage is maximal at the lesion epicenter, or the center of the injury site, and is directly related to the severity of the initial injury (Pavel). Early identification of an IPH is crucial to maximizing the chances of the individual's survival, with factors such as blood pressure, reversal of associated coagulopathy (correcting the condition causing the impaired formation of blood clots), intensive care, and identifying the secondary etiologies (causes of condition) (Gross, Jankowitz, & Freidlander, 2019).

A serious consequence of TBIs is meningitis, which is a fatal disease caused by the inflammation of the membranes surrounding the brain and spinal cord (WHO, 2023). The triple-layered membrane surrounding the brain and spinal cord is called the meninges, and the three layers from outermost to innermost are called the dura, arachnoid, and pia mater as previously mentioned in the paper (Ghannam & Kharazi, 2023). Although bacterial meningitis is more prevalent in meningitis patients (WHO, 2023) and can be a complication of TBIs (Santos et al., 2011), posttraumatic meningitis is a type of infection that occurs because of a TBI, and negatively affects the outcome of the patient severely (Katayama et al., 2021). More specifically, posttraumatic meningitis can result from a basilar fracture in the head and a condition called cerebrospinal fluid fistula (CSF fistula), which is an abnormal leakage of cerebrospinal fluid (Katayama et al., 2021). Posttraumatic meningitis significantly increases the chance of mortality with reports stating that mortality rates range from 29% to 57.9% when meningitis is involved (Katayama et al., 2021). Moreover, it can be difficult to identify the cause of meningitis because clinical signs are nonspecific and lack pathognomonic laboratory markers (characteristics of a disease) (La Russa et al., 2020). Additionally, these markers increase with a long latency, which hinders a prompt diagnosis (La Russa et al., 2020). Overall, this disease has several links to TBIs and remains a huge topic in the realm of neurological diseases.

Diffuse Axonal Injury and Brain Herniation

Another type of TBI is diffuse axonal injuries or DAIs. DAIs are severe TBIs characterized by delayed axonal disconnection, which occurs because of damage to axons (Mu et al., 2019). However, the true incidence of DAI is unknown (Mesfin, Gupta, Shapshak, & Taylor, 2023). Estimates state that 10% of all TBI patients will have an incidence of DAI, and 25% of persons with DAI will result in death (Mesfin, Gupta, Shapshak, & Taylor, 2023). This makes DAIs one of the most common forms of TBIs (Mu et al., 2019) and can occur from rapid acceleration/deceleration. To provide further clarification, axons are elongated portions of the neuron that send electrical impulses and projects to synapses with dendrites or cell bodies of other neurons (Muzio, 2022).

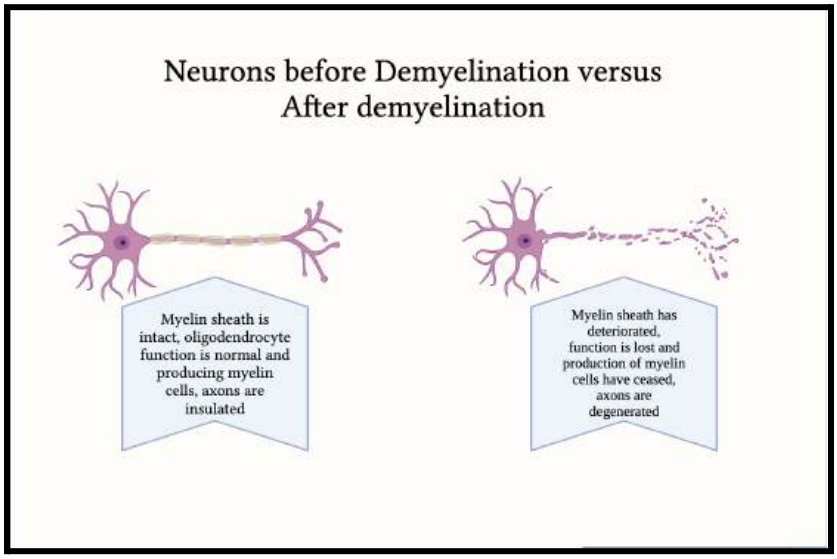


Figure 2: A comparison of two neurons: a healthy, myelinated neuron and a demyelinated neuron. The right neuron depicts axonal demyelination, a hallmark of diffuse axonal injury. Created with BioRender.

Furthermore, axons are coated by a fatty protein called the myelin sheath, which is produced by glial cells called oligodendrocytes in the central nervous system and plays a huge role in axon maintenance (Mu et al., 2019). One can make an analogy that the myelin sheath is like the rubber insulation used in electrical wires. However, in stretch injuries to optic nerve fibers, myelin dislocations occur within one to two hours after injury and the myelin sheath and oligodendrocytes face injury and damage, which facilitates axon degeneration (Mu et al., 2019). The process in which the myelin sheath degenerates is called demyelination. Because of the results found in stretch injuries to optic nerve fibers, there has been a strong link between myelin damage and the pathophysiological processes that occur during a DAI (Mu et al., 2019). The relationship between axons and the brain is critical in understanding the true scope of damage that a DAI can inflict on an individual. Because axons are responsible for sending electrical impulses to other neurons (Mu et al., 2019), this ensures that the brain interacts with day-to-day life by allowing it to coordinate behavior, sensation, thoughts, and emotion (Trafton, 2018). In light of all this, there is a spectrum of neurological dysfunction that occurs because of DAI (Mesfin, Gupta, Shapshak, & Taylor, 2023). This can vary from clinically insignificant to

rendering the individual in a comatose state (Mesfin, Gupta, Shapshak, & Taylor, 2023). To elaborate further, the axonal damage from DAI in the cerebral white matter is associated with commas. Moreover, an increase in DAI lesions is associated with a higher severity of cognitive impairment, as discovered by magnetic resonance imaging (MRI). Overall, DAI is a severe form of TBI and can demonstrate a myriad of neurological complications at a level more advanced than concussions, skull fractures, or even intracranial hemorrhaging. However, some TBIs are more life-threatening and like a DAI, require serious aid for fear of death.

Traumatic brain herniations are a devastating form of severe TBI that requires immediate attention and an early diagnosis (Munakomi & Das, 2023). During a herniation, the pressure inside the skull increases and cerebrospinal fluid is displaced within the skull (Munakomi & Das, 2023). These failures within the skull, if left without treatment, will result in a series of pathological cascades which can lead to respiratory arrest and certain death (Munakomi & Das, 2023). The urgency that this TBI must be treated as soon as possible cannot be stressed enough, with the alternative name being called brain code (Munakomi & Das, 2023), for "code" is a term used in medical settings to indicate immediate medical attention. To further know about the definition of brain herniation, we must familiarize ourselves with the Monro-Kellie doctrine.

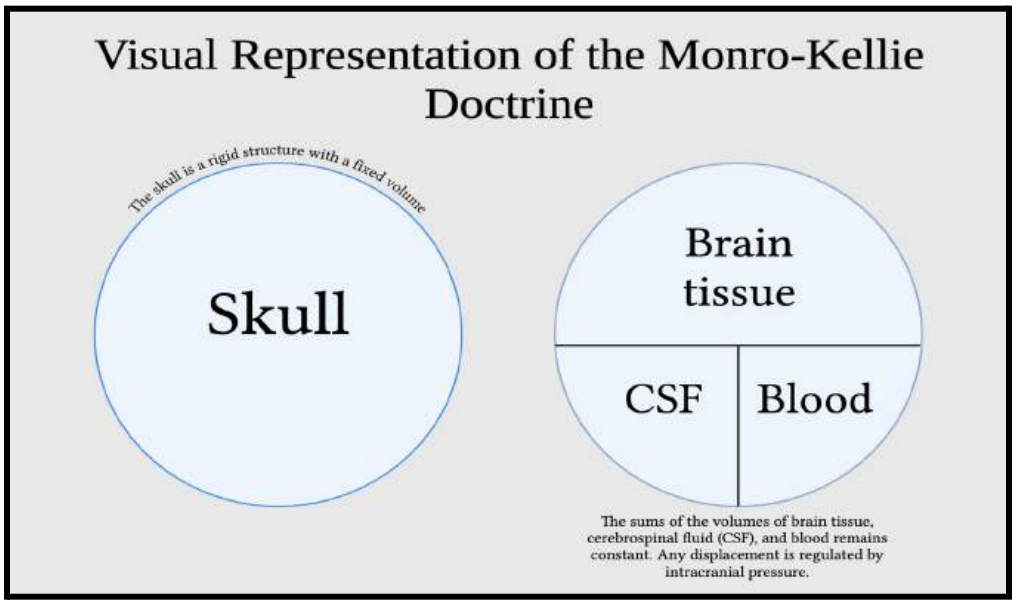


Figure 3: Simplistic model of the balance of CSF, brain tissue, and blood as modeled by the Monro-Kellie Doctrine. Created using Biorender.

This doctrine states that the sums of volumes of the brain, cerebrospinal fluid, and intracranial blood remain constant, and any changes to one of the volumes affect the other two volumes (Mokri, 2001). Intracranial pressure (ICP) plays a role as a regulatory factor: when one of the three factors of the doctrine fluctuates in volume, the ICP regulates the displaced factors to maintain a constant volume. During a brain herniation, ICP dramatically increases, and all compensatory mechanisms are overridden, thus, causing a severe violation of the Monro-Kellie

doctrine (Munakomi & Das, 2023). Given our knowledge of the Monro-Kellie doctrine, brain herniation is caused by a significant increase in ICP, which leads to displacement of either cerebrospinal fluid, intracranial blood, or brain tissue. There are a myriad of factors that may lead to an increase in ICP, such as intracerebral hemorrhaging, tumors, malignant infarctions (tissue death), and CSF over drainage, to name a few (Munakomi & Das, 2023). There are five types of brain herniations, all of which are severe: surfacing, transtentorial, central, tonsillar, and upward herniations (Munakomi & Das, 2023).

Foremost, subfalcine herniations are the most common type of herniation, and they occur when brain tissue is displaced under the falx cerebri (Kostecki, 2023), which is the sickle-shaped fold of the dura mater and separates the right and left cerebral hemispheres (Bair & Munakomi, 2023).

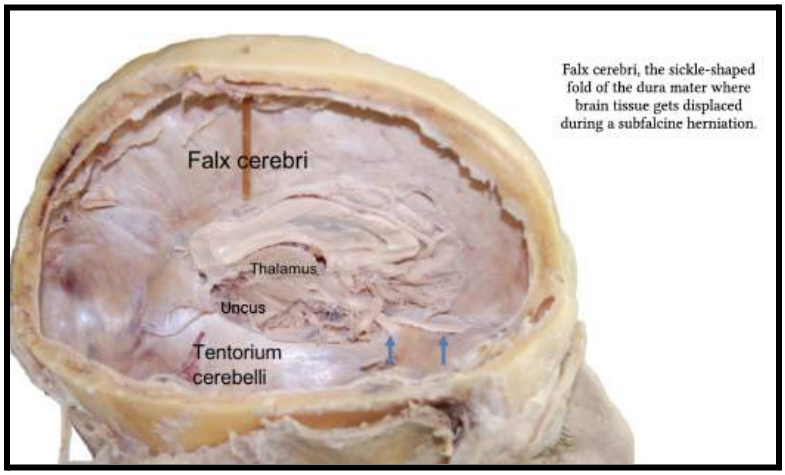


Figure 4: A right lateral view into the location of the falx cerebri in an unaffected brain to fully understand the scope of subfalcine herniation. The labeled structures demonstrate key components of the brain's protective and structural anatomy, including the dural folds and central brain regions. Created using Biorender.

A subfalcine herniation is a secondary intracranial injury (Kostecki, 2023), often resulting from primary factors such as increased brain volume, hemorrhaging, CSF volume changing, heart failure, and rapid decompression of a hematoma (Kostecki, 2023). Initially, a subfalcine herniation may not cause severe symptoms, with symptoms as benign as a headache being the first sign of a herniation (Kostecki, 2023). Additionally, it is not uncommon to experience an altered mental state, nausea, and vomiting (Kostecki, 2023). Ultimately, individuals suffering from a subfalcine herniation need immediate medical care, and since initial symptoms are subtle, it is necessary to remain in an open line of communication with the patient (Kostecki, 2023).

Furthermore, transtentorial herniations occur when brain tissue moves from one intracranial compartment to the other (Knight, 2023) because of rising ICP rates (Decker, 2023). These intracranial departments consist of the supratentorial space (PDQ Pediatric Treatment Editorial Board, 2023), the infratentorial space (Jeong, 2018), and the ventricular space (Shenoy,

2023). This type of herniation is life-threatening and requires immediate medical attention and may be reversible with emergent surgical intervention and medical management. The incidence of transtentorial herniations is poorly documented, however, they occur as a secondary injury to traumatic brain injuries (Knight, 2023; Decker, 2023). Due to the nature of brain tissue displacement from one intracranial compartment to another, a transtentorial herniation also violates the previously stated Monro-Kellie Doctrine (Knight, 2023). Furthermore, a transtentorial herniation is any parenchymal herniation through the tentorial notch, and there are multiple types of ways a herniation can occur through the notch, such as uncal, central, and upward herniations (Knight, 2023).

Uncal herniations are associated with the uncal portion of the temporal lobe (Knight, 2023). Moreover, the uncus is a structure at the anterior medial aspect of the parahippocampal gyrus in the medial temporal region of the brain (Decker, 2023). To clarify, the uncus is also the structure affected during an uncal herniation. Compartment-wise, this herniation takes place in the supratentorial compartment, which is connected to the subtentorial compartment in the brain (Decker, 2023). Furthermore, central herniations differ from uncal herniations: the pressure shift from an uncal herniation causes medial and downward herniation of the uncus, and a central herniation is due to direct pressure on the affected region (Knight, 2023).

Central herniation can cause rare yet fatal medical complications. For example, central herniation can compress the pituitary stalk, which acts as a conduit for hormonal regulation and nerve signals (Knight, 2023). As a result, there is a lack of antidiuretic hormone (ADH) (Knight, 2023), which plays a key role in regulating water levels in the body (Cuzzo, 2023), If ADH is minimal or not present, the individual is then afflicted with diabetes insipidus, or water diabetes; this rare condition is characterized by constant thirst and urination, resulting in the body losing water (Knight, 2023). This may lead to necrosis of the midbrain and the formation of slit-like hemorrhages called Duret hemorrhages (Knight, 2023). Lastly, upward herniations, also known as upward transtentorial herniation, occur when an infratentorial mass, like a tumor or a cerebellar hemorrhage, compresses the brain stem, kinking it and causing patchy brain stem ischemia (Maiese, 2024). Also known as a brain stem stroke, patchy brain stem ischemia is one of the most lethal types of strokes, with ischemic brainstem strokes and hemorrhagic strokes making up many morbidities and mortality in the United States (Gowda, 2024). Additionally, these strokes make up 10-15% of all strokes (Gowda, 2024). Ultimately, transtentorial herniations are a multifaceted area of traumatic brain injury, in which multiple herniations fall under this category.

A tonsillar herniation is the movement of brain tissue into one intracranial compartment, specifically the movement of the cerebellar tonsils through the foramen magnum (Knight, 2023). The nature of this herniation is quite severe, with the individual requiring exigent medical care and surgical intervention (Knight, 2023). Specifically, the pressure exerted by the tonsils onto the magnum causes the medulla to compress against the clivus/odontoid process, in a process called “coning” (Knight, 2023). This name comes from the fact that brain tissue is squeezed down the foramen like a cone (Knight, 2023). Sadly, if the pressure exerted onto the medulla progresses, it

will trigger Cushing's reflex—a physiological response to elevated ICP characterized by hypertension, bradycardia, and irregular respiration (Dinallo, 2023)—which, if untreated, can lead to death (Knight, 2023). This herniation can be caused by numerous other traumatic injuries, such as intracerebral hemorrhaging, subarachnoid hemorrhage, and DAI (Knight, 2023). As one may have noticed, there is a pattern for these types of severe traumatic injuries to violate the Monro-Kellie Doctrine, making it a reliable way to interpret TBI. Tonsillar herniations are no exception, for the Monro-Kellie Doctrine is also violated here (Knight, 2023). One notable example of tonsillar herniation is cerebellar herniation, which can cause rapid herniations through the foramen magnum and cause the compression of the medulla. Ultimately, traumatic brain herniations are a multifaceted area of traumatic brain injury, in which multiple herniations fall under this category. It is a fatal, dangerous cause of morbidity and mortality in the United States.

Treatments

There are currently no FDA-approved treatments or devices that can assess and diagnose TBI (FDA, 2021). However, this doesn't stop modern medicine from trying new therapies to prevent and relieve TBI. For example, amantadine and memantine are drugs commonly used to slow down neurodegenerative disorders such as Parkinson's (Laplante & St. George, 2014) and Alzheimer's disease (Kuns, Rosani, Patel, & Varghese, 2024), respectively. Even though these drugs are also commonly used in acute rehabilitation settings following TBI, a lack of approval exists for neuro-recovery by the FDA ((Ma & Zafonte, 2020). Unfortunately, the research done for testing amantadine and memantine for TBI recovery, although positive, was tested in small population sizes (Ma & Zafonte, 2020). Additionally, the populations that are being tested with these drugs may also have a bias (Ma & Zafonte, 2020). Moreover, the responses, or side effects given by amantadine and memantine are inconsistent, making it even harder to mark the two drugs as effective for TBI and remain an elusive result (Ma & Zafonte, 2020).

The ways that hospitals have looked at neurocritical care and monitoring have changed, and doctors and surgeons look to redefine the type of neurocritical care and monitoring that occurs to treat and save a patient with TBI. Firstly, the concept of primary and secondary injuries became well-known around twenty-five years ago, with the simple recognition that around the time of trauma, additional medical complications such as hypotension and hypoxia can arise, which can exacerbate TBI (Khellaf, Khan, & Helmy, 2019).

One example of neurocritical care doctors use is intracranial pressure monitoring, or ICP monitoring (Khellaf, Khan, & Helmy, 2019). As previously stressed in this review, ICP is a critical component of TBI and is a widely accepted treatment among the TBI community (Khellaf, Khan, & Helmy, 2019). However, a recent randomized control trial monitoring of ICP didn't show any benefit (Khellaf, Khan, & Helmy, 2019). This trial was severely criticized for two key details: the trial was carried out in units where there was no previous monitoring of ICP, and both groups of patients tested had aggressive ICP therapies irrespective (Khellaf, Khan, & Helmy, 2019).

Another example of neurocritical care is multimodality neuromonitoring, or MNM (Lazaridis & Foreman, 2023). The purpose of MNM is to integrate and interpret multiple sources of information so doctors can remain equipped with the development of severe TBIs (Lazaridis & Foreman, 2023). Even though there isn't a set number of modalities or treatments, there is a multitude of measurements used, such as compartmental (e.g., intracranial) and perfusion pressures (pressure required to ensure adequate blood flow to the brain), tissue oxygenation and metabolism, pressure autoregulation, electrophysiology (Lazaridis & Foreman, 2023). Furthermore, there is a growing recognition that the use of integrated, complementary information may be better suited to individualized care rather than dependence on a multipurpose approach. Ultimately, although there aren't any FDA-approved treatments or devices for traumatic brain injuries, efficacious therapies are being used despite the lack of FDA approval and the growth of treatments in the TBI community continues to gain research and more information.

Neurological Disorders

To start the final section of the review, we stray away from traumatic-induced injuries and focus on neurological disorders. To understand and gain a wider scope of neuroscience, one must not only take traumatic-induced external conditions into account but also genetic or systematic-induced internal conditions. Out of the numerous types of neurological disorders that exist in the medical world, I have singled out five neurological disorders that are of great importance and are worth knowing about. This subsection will be divided into five parts entailing a different disorder: Alzheimer's disease, multiple sclerosis, Parkinson's disease, Huntington's disease, and autism.

Alzheimer's Disease

Alzheimer's disease is a major cause of dementia and is a progressive neurodegenerative disorder that induces memory loss, and difficulties with thinking, language, and problem-solving

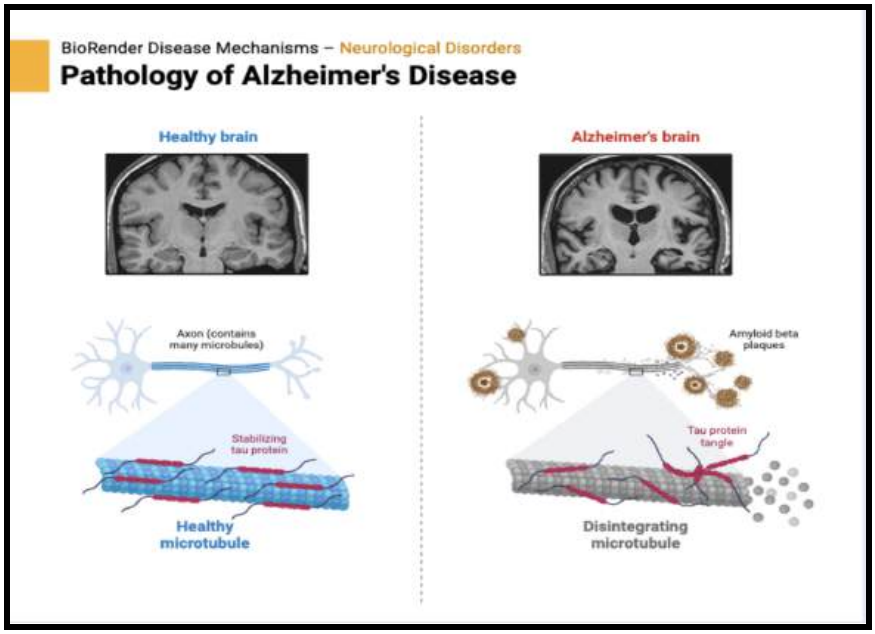


Figure 5: An insight into neurons unaffected by Alzheimer's disease versus neurons affected by Alzheimer's disease, with amyloid-beta plaques and tau protein tangles accumulating in the microtubule, causing it to disintegrate. Created using Biorender.

abilities (Khan, Barve, & Kumar, 2020). Furthermore, Alzheimer's disease causes the brain cells in a patient to degenerate. Moreover, there are a multitude of risk factors involved in this disorder, such as age, genetics, head injuries, vascular diseases, infections, and environmental factors (Breijyeh & Karaman, 2020). Alzheimer's disease is considered a multifactorial disease: the cause of the disease is dependent on several factors (Breijyeh & Karaman, 2020). There were two hypotheses proposed as key causes of Alzheimer's: cholinergic and amyloid hypotheses (Breijyeh & Karaman, 2020). Alzheimer's disease, which is also the most common type of dementia, can be characterized as the result of amyloid-beta peptide (Aβ) accumulation in the medial temporal lobe and the neocortical structures, which are the most affected areas of the brain (Breijyeh & Karaman, 2020). Alois Alzheimer, the man the disease is named for, noticed the presence of amyloid plaques (because of accumulation) and an enormous loss of neurons in the patient (Breijyeh & Karaman, 2020). Furthermore, the patient suffered from memory loss and a change in personality before dying, for which Alzheimer declared this condition to be a severe disease of the cerebral cortex, the part of your brain responsible for processing information (Breijyeh & Karaman, 2020). Currently, the only way to definitively diagnose Alzheimer's would be to perform an autopsy on the person's brain, but because a doctor cannot perform an autopsy on a living person, diagnostic tests are done to accurately determine the disease instead (Khan, Barve, & Kumar, 2020). Ultimately, Alzheimer's disease is a multifactorial progressive neurodegenerative disorder that remains a main cause of dementia and can have severe life changes on the afflicted.

Multiple Sclerosis

Multiple sclerosis is a potentially progressive autoimmune neurologic disorder of the central nervous system, which results from an autoimmune attack on the central nervous system white matter (Cotsapas, Mitrovic, & Hafler, 2018).

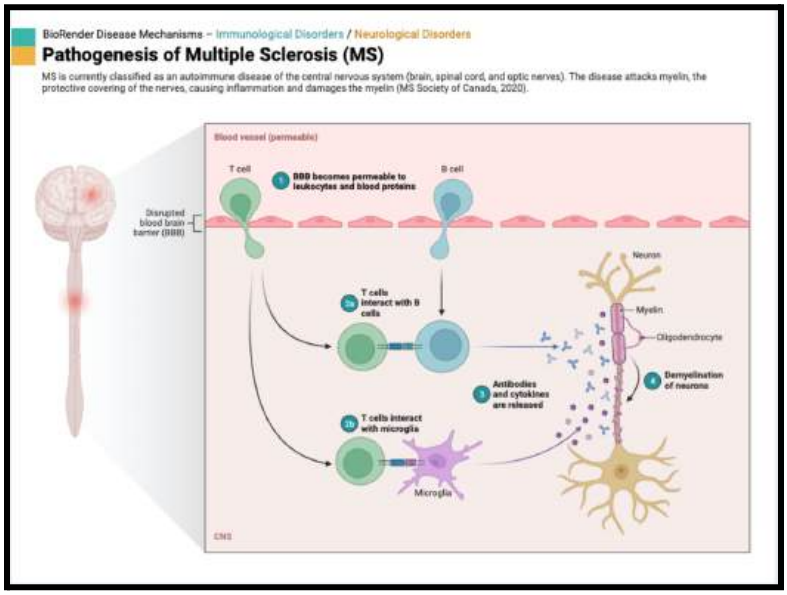


Figure 6: A molecular view into the pathogenesis of multiple sclerosis. To simplify the diagram, the antibodies and cytokines are released, attacking the neuron and resulting in demyelination. Created using Biorender by Ruslan Medzhitov, Akiko Iwasaki, Jung-Hee Lee.

Essentially, multiple sclerosis causes the body to attack itself. There are two disease courses: relapse remitting MS (RRMS) and progressive MS. For the sake of the review, we will cover progressive MS.Currently, there is no known cure for multiple sclerosis, and it is a leading cause of neurologic symptoms in young adults (Cotsapas, Mitrovic, & Hafler, 2018). Furthermore, the role of genetic inheritance has been proven to play a role in multiple sclerosis (Patsopoulos, 2018). Additionally, 15-20% of multiple sclerosis patients have had a family history of the disease (Patsopoulos, 2018). To further solidify the role of genetic inheritance, it is more likely for an offspring to get multiple sclerosis if both parents have multiple sclerosis (Patsopoulos, 2018). At the cellular level, the myelin sheath is demyelinated because of an inflammatory attack, causing neuro-axonal damage (Lemus, Warrington, & Rodriguez, 2017). Additionally, there is no cure for multiple sclerosis because there is no effective treatment to halt or reverse neuro-axonal damage or promote remyelination (Lemus, Warrington, & Rodriguez, 2017). Additionally, multiple sclerosis is an organ-specific disease, targeting the brain and spinal cord, with immune-mediated myelin destruction (Lemus, Warrington, & Rodriguez, 2017). As previously mentioned in our discussion, demyelination, and axon degeneration are deadly and can cause some serious medical repercussions to the body. Ultimately, multiple sclerosis is a deadly neurodegenerative disease and it can have serious medical repercussions for anyone afflicted with the deadly disease.

Parkinson's Disease

Parkinson's disease is one of the most common chronic neurodegenerative diseases in the world and is characterized by both motor and non-motor symptoms (Beitz, 2014). Parkinson's disease is also an idiopathic disease of the nervous system, which essentially means the cause is unknown (Beitz, 2014). Some symptoms of Parkinson's disease are rest tremors, or involuntary movement of a body part (Chen, Hopfner, Becktepe, & Deuschl, 2017), rigidity, and bradykinesia (Beitz, 2014). Bradykinesia is the generalized slowing of movement (Zafar & Yaddanapudi, 2023). There are also nonmotor symptoms associated with Parkinson's disease, such as cognitive changes, behavioral and neuropsychiatric changes to the nervous system, sleep issues, and sensory problems (Beitz, 2014). This disease most commonly occurs in older people, but younger persons can also develop Parkinson's disease (Beitz, 2014). Despite its idiopathic status, doctors know it is associated with the loss or degeneration of the dopaminergic (dopamine-producing) neurons in the substantia nigra and the development of Lewy bodies in these dopaminergic neurons (Beitz, 2014). Furthermore, the substantia nigra is the part of the brain that controls motor movement which is a part of the basal ganglia (Sonne, Reddy, & Beato, 2022). Moreover, Lewy bodies are filamentous, cytoplasmic aggregations of protein that build up in dopaminergic neurons, resulting in brain damage (Meredith, Halliday, & Totterdell, 2004).

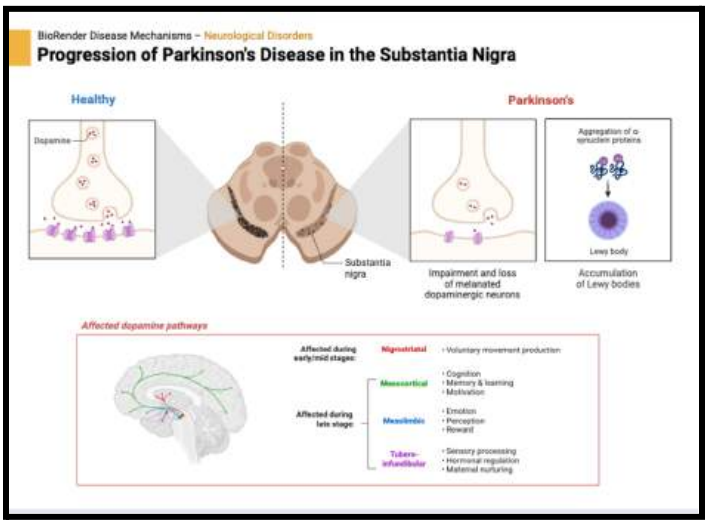


Figure 7: This figure represents the loss of dopaminergic neurons and the dysfunction of the substantia nigra within dopamine pathways. Created using Biorender by Allison Kufta and Sally Kim.

There are a multitude of risk factors that are associated with Parkinson's disease. The most common risk factor is age, for this disease most commonly occurs in older persons, but younger persons can also develop Parkinson's disease (Beitz, 2014). Moreover, two other causes are crucial, such as pesticide exposure and family history (Beitz, 2014). The myriads of complications that are caused by Parkinson's disease are bound to take a toll on the individual's daily life. For instance, persons with the disease tend to suffer from depression, excessive daytime sleepiness, anxiety, and apathy (Al-Khammash et al., 2023). Ultimately, Parkinson's

disease is a deadly neurodegenerative disease that affects millions of unfortunate individuals per year, and it can severely decrease the quality of living for the afflicted.

Autism Spectrum Disorder

Autism spectrum disorder, or ASD, is a behaviorally defined neurodevelopmental disorder (Wang, Wang, Wu, Wang, & Sun, 2023). It is characterized by a collection of genetically variant neurodevelopmental disorders that manifest as early social intercourse dysfunction and impaired repetitive behaviors and interests (Wang, Wang, Wu, Wang, & Sun, 2023). The prevalence of autism is more common in boys than girls, and 1 out of 44 children, or 2.27% are diagnosed (Wang, Wang, Wu, Wang, & Sun, 2023). Along with ASD, co-occurring conditions can also show, such as depression, epilepsy, anxiety, and attention deficit hyperactivity disorder (ADHD). Persons with autism have atypical cognitive deficits, such as impaired social cognition and awareness, executive dysfunction, and atypical perception and information processing (Wang, Wang, Wu, Wang, & Sun, 2023). The genetic heritability of ASD varies from 40-90%, with estimates of 50% genetic liability (Sanchack & Thomas, 2016). The severity of autism can vary, and it is determined by the symptoms that the individual showcases during diagnosis.

Autism Severity Levels

Severity Level	Social Communication	Restricted and repetitive behaviors
Level 1. Requiring support	(i.e., Without supports in place, deficits in social communication cause noticeable impairments. Has difficulty initiating social interactions and demonstrates clear examples of atypical or unsuccessful responses to social overtures of others. May appear to have decreased interest in social interactions.)	(i.e., Rituals and repetitive behaviors [RRBs] cause significant interference with functioning in one or more contexts. Resists attempts by others to interrupt RRBs or to be redirected from fixated interest.)
Level 2: Requiring substantial support	(i.e., Marked deficits in verbal and nonverbal social communication skills; social impairments apparent even with supports in place; limited initiation of social interactions and reduced or abnormal response to social overtures from others.)	(i.e., RRBs and/or preoccupations and/or fixated interests appear frequently enough to be obvious to the casual observer and interfere with functioning in a variety of contexts. Distress or frustration is apparent when RRBs are interrupted; difficult to redirect from fixated interest.)
Level 3: Requiring very substantial support	(i.e., Severe deficits in verbal and nonverbal social communication skills cause severe impairments in functioning; very limited initiation of social interactions and minimal response to social overtures from others.)	(i.e., Preoccupations, fixed rituals and/or repetitive behaviors markedly interfere with functioning in all spheres. Marked distress when rituals or routines are interrupted; very difficult to redirect from fixated interest or returns to it quickly.)

Table 1: Adapted from the American Psychiatric Association - APA. The table above represents the signs and symptoms of each level of autism severity, from Level 1 to Level 3.

There is no general agreement on what constitutes an appropriate measure of quality of life for autistic individuals (Øverland et al., 2022). Autistic individuals may have a different perception of an ideal quality of life than neurotypical individuals (Øverland et al., 2022). There are autism-specific qualities of life, such as autism burnout. Essentially, autism burnout is characterized as a severe condition of fatigue added by social withdrawal, cognitive dysfunctions, and exacerbation of autism traits (Øverland et al., 2022). Moreover, it differs from burnout suffered by depression and job-related stress (Øverland et al., 2022). Ultimately, ASD is a neurodevelopmental disorder that has a huge social and mental impact on the individual, based on the severity of the diagnosis, and it is important to remember that autistic individuals face difficulties unknown to neurotypical individuals, so actions must be taken to remove the stigma of autism in society.

Huntington's Disease

Huntington's disease is a dominantly inherited neurodegenerative disease caused by repeated CAG segments on the *huntingtin (HTT)* gene and is characterized by progressive motor, cognitive, and neuropsychiatric decline (Pengo & Squitieri, 2024).

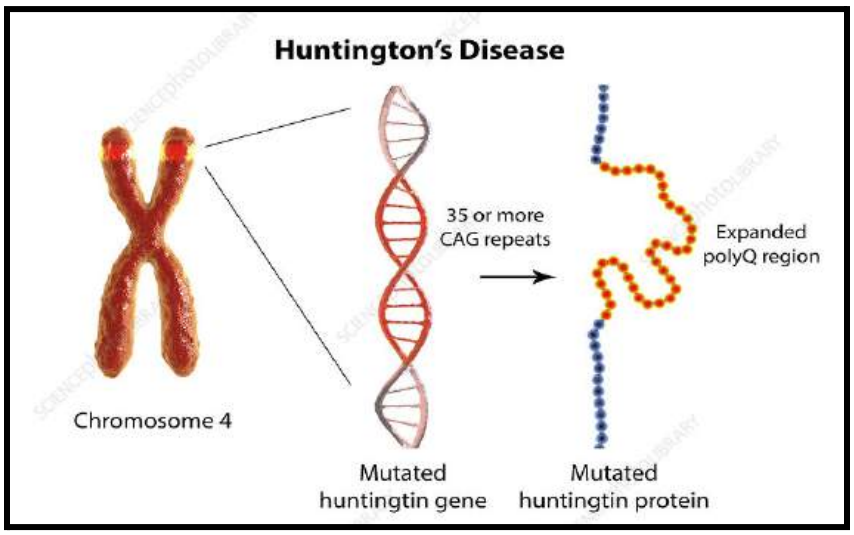


Figure 8: The diagram illustrates the mutated huntingtin gene because of multiple CAG repeats on chromosome 4. Created by Kateryna Kon.

Furthermore, the disease is also characterized by shrinkage of the brain, loss of efferent medium spiny neurons (MSNs), and degeneration of the striatum, which controls motor skills (Jimenez-Sanchez, Licitra, Underwood, & Rubinsztein, 2017). A regionally specific thinning of the cortical ribbon has also been found to be present in patients (Jimenez-Sanchez, Licitra, Underwood, & Rubinsztein, 2017). Because of this, the disease spreads throughout the brain, more specifically, the posterior and anterior cortical regions (Jimenez-Sanchez, Licitra, Underwood, & Rubinsztein, 2017). As mentioned before, the expression of the mutant *HTT* gene contributes to the development of Huntington's disease. There is a huge genetic factor in play when it comes to the development of Huntington's disease. Huntingtin contains a polyglutamine tract encoded by uninterrupted CAG repeats in the first exon of *HTT* (Jimenez-Sanchez, Licitra,

Underwood, & Rubinsztein, 2017). Exons are the coding regions in genes, while CAG represents three of the four nucleotide bases in DNA: cytosine, adenine, guanine, and thymine. Moreover, if the repeated CAG expansions in *HTT* are immense, the age of onset for Huntington's disease decreases, which essentially means that individuals with large expansions of the DNA segment will develop Huntington's disease at an earlier age (Jimenez-Sanchez, Licitra, Underwood, & Rubinsztein, 2017). The quality of life for persons with Huntington's disease is decreased compared to neurotypical individuals, with patients showing cognitive and behavioral issues, changes to mood and temperament, and the hallmark of Huntington's disease: motor dysfunction (Eddy & Rickards, 2022). Some examples of social cognition deficiency, as expressed by patients with Huntington's disease, include not picking up others' mental states, struggling to understand emotion, facial expressions, and posture, and not understanding social interactions (Eddy & Rickards, 2022). Despite this research, there is not much scientific research about the quality of life for people with Huntington's disease (Eddy & Rickards, 2022). Ultimately, Huntington's disease is a neurodegenerative disease that is either inherited from your parents or caused because of mutant *HTT* genes, and this disease can cause severe damage to the brain and impair motor functions, like many other diseases of its kind like Parkinson's disease.

Conclusion

Traumatic brain injuries (TBI) occur by withstanding trauma to the head from a shock, jolt, or blow. Additionally, this paper is a comprehensive review of studies that ranged from multiple sources regarding how TBIs affects the human mind mentally, cognitively, and physically. TBIs can range from multiple degrees of severity from mild TBIs (mTBIs), moderate TBIs, and severe TBIs. This paper rigorously analyzes the multiple types of TBIs in each degree of severity and how the implications of such affect the human mind, such as damages to vital brain regions, traumatic-induced infections, internal bleeding damage, and damage done to the brain at the molecular/axonal level. This paper attempts to bridge key gaps in TBI knowledge, like the cognitive and mental impact of TBIs, long-term effects of TBI, and the current state of research regarding TBI. Future TBI research should focus on advancing diagnostic technologies like biomarkers and imaging, exploring personalized treatments such as stem cell therapy and precision medicine, and conducting long-term studies on neurodegeneration and rehabilitation outcomes. Additionally, addressing underserved populations, developing innovative protective strategies, and fostering cross-disciplinary collaboration can significantly enhance prevention, treatment, and recovery efforts.Moreover, the paper also delves into research and papers about certain well-known neurological diseases, like Huntington's disease, Alzheimer's disease, Parkinson's disease, autism spectrum disorder, and multiple sclerosis. Ultimately, the main goal of this review is to compile essential information about TBI into a single paper and help contribute to developing therapies and most importantly, raising awareness about the multiple types of TBI and neurological conditions that millions of people are diagnosed with.

Works Cited

Al-Khammash, Noora, et al. "Quality of Life in Patients with Parkinson's Disease: A Cross-Sectional Study." Cureus, U.S. National Library of Medicine, 20 Jan. 2023, www.ncbi.nlm.nih.gov/pmc/articles/PMC9941031/.

Alao, Titilola, et al. "Penetrating Head Trauma." StatPearls [Internet]., U.S. National Library of Medicine, 30 Jan. 2024, www.ncbi.nlm.nih.gov/books/NBK459254/.

Bair, Michael M, and Sunil Munakomi. "Neuroanatomy, Falx Cerebri." StatPearls [Internet]., U.S. National Library of Medicine, 24 July 2023, www.ncbi.nlm.nih.gov/books/NBK545304/#:~:text=The%20fax%20cerebri%20is%20a,arachnoid%20mater%2C%20and%20dura%20mater.

Beitz, Janice M. "Parkinson's Disease: A Review." Frontiers in Bioscience (Scholar Edition), U.S. National Library of Medicine, 1 Jan. 2014, pubmed.ncbi.nlm.nih.gov/24389262/.

Bell, Daniel J. "Skull Sutures: Radiology Reference Article." Radiopaedia, Radiopaedia.org, 14 Apr. 2024, radiopaedia.org/articles/skull-sutures-1?lang=us.

Bielanin, John P, et al. "An Overview of Mild Traumatic Brain Injuries and Emerging Therapeutic Targets." Neurochemistry International, Pergamon, 9 Dec. 2023, www.sciencedirect.com/science/article/pii/S0197018623001833?via%3Dihub.

Breijyeh, Zeinab, and Rafik Karaman. "Comprehensive Review on Alzheimer's Disease: Causes and Treatment." Molecules (Basel, Switzerland), U.S. National Library of Medicine, 8 Dec. 2020, pubmed.ncbi.nlm.nih.gov/33302541/.

Cadman, Bethany. "Skull Fracture: Types, Symptoms, and Long-Term Effects." Edited by Seunggu Han, Medical News Today, MediLexicon International, 24 July 2024, www.medicalnewstoday.com/articles/322871.

Capizzi, Allison, et al. "Traumatic Brain Injury: An Overview of Epidemiology, Pathophysiology, and Medical Management." Medical Clinics of North America, Elsevier, 6 Feb. 2020, www.sciencedirect.com/science/article/abs/pii/S0025712519301294?via%3Dihub.

Chen, Wei, et al. "Rest Tremor Revisited: Parkinson's Disease and Other Disorders." Translational Neurodegeneration, U.S. National Library of Medicine, 16 June 2017, www.ncbi.nlm.nih.gov/pmc/articles/PMC5472969/.

Cotsapas, Chris, et al. "Multiple Sclerosis." Handbook of Clinical Neurology, Elsevier, 22 Feb. 2018, www.sciencedirect.com/science/article/abs/pii/B9780444640765000466?via%3Dihub.

Cuzzo, Brian, et al. "Physiology, Vasopressin." StatPearls [Internet]., U.S. National Library of Medicine, 14 Aug. 2023, www.ncbi.nlm.nih.gov/books/NBK526069/.

Decker, Rebecca, and Anthony L Pearson-Shaver. "UNCAL Herniation." StatPearls [Internet]., U.S. National Library of Medicine, 8 Aug. 2023, www.ncbi.nlm.nih.gov/books/NBK537108/#:~:text=Uncal%20herniation%20occurs%20when%20rising,adaptive%20mechanisms%20for%20intracranial%20compliance.

Dinallo, Sean, and Muhammad Waseem. “Cushing Reflex.” StatPearls [Internet]., U.S. National Library of Medicine, 20 Mar. 2023, www.ncbi.nlm.nih.gov/books/NBK549801/.

Eddy, Clare M, and Hugh Rickards. “Social Cognition and Quality of Life in Huntington’s Disease.” Frontiers in Psychiatry, U.S. National Library of Medicine, 24 Aug. 2022, www.ncbi.nlm.nih.gov/pmc/articles/PMC9449535/.

Ghannam, Jack Y, and Khalid A. Al Kharazi. “Neuroanatomy, Cranial Meninges.” StatPearls [Internet]., U.S. National Library of Medicine, 24 July 2023, www.ncbi.nlm.nih.gov/books/NBK539882/#:~:text=The%20 arachnoid%20sits%20between%20the,cortical%20 sulci%20but%20 bridges%20theme.

Gowda, Supreeth N, et al. “Brainstem Stroke.” StatPearls [Internet]., U.S. National Library of Medicine, 25 Feb. 2024, www.ncbi.nlm.nih.gov/books/NBK560896/.

Gross, Bradley A, et al. “Cerebral Intraparenchymal Hemorrhage.” JAMA, JAMA Network, 2 Apr. 2019, jamanetwork.com/journals/jama/article-abstract/2729375.

Hyder, Adnan A, et al. “The Impact of Traumatic Brain Injuries: A Global Perspective.” IOS.Press, NeuroRehabilitation, 2007, content.iospress.com/download/neurorehabilitation/nre00374?id=neurorehabilitation/nre00374.

Jeong, Seok-Hoo, et al. “Endoscopic Botulinum Toxin Injection for Treatment of Pharyngeal Dysphagia in Patients with Cricopharyngeal Dysfunction.” Taylor and Francis, Scandinavian Journal of Gastroenterology, 24 Oct. 2018, www.tandfonline.com/doi/10.1080/00365521.2018.1506820?_ga=2.196256815.2007054471.1722554302-254167365.1722554302&_gl=1*1st hs4e*_ga*MjU0MTY3MzY1LjE3MjI1NTQzMDI.*_ga_0HYE8YG0M6*MTcyMjU1NDMwMy4xLjAuMTcyMjU1NDMwMy4wLjAuMA..*_gcl_au*MjAxMDQxNzQ1MS4xNzIyNTU0MzA0.

Jimenez-Sanchez, Maria, et al. “Huntington’s Disease: Mechanisms of Pathogenesis and Therapeutic Strategies.” Cold Spring Harbor Perspectives in Medicine, U.S. National Library of Medicine, 5 July 2017, www.ncbi.nlm.nih.gov/pmc/articles/PMC5495055/.

Kairys, Norah, et al. “Acute Subarachnoid Hemorrhage.” StatPearls [Internet]., U.S. National Library of Medicine, 10 Oct. 2022, www.ncbi.nlm.nih.gov/books/NBK518975/#:~:text=Subarachnoid%20hemorrhages%20are%20true%20emergencies,pia%20mater%20surrounding%20the%20brain.

Katayama, Yusuke, et al. “Factors Associated with Posttraumatic Meningitis among Traumatic Head Injury Patients: A Nationwide Study in Japan.” European Journal of Trauma and Emergency Surgery : Official Publication of the European Trauma Society, U.S. National Library of Medicine, 3 Sept. 2021, www.ncbi.nlm.nih.gov/pmc/articles/PMC7851005/.

Khairat, Ali, and Muhammad Waseem. “Epidural Hematoma.” StatPearls [Internet]., U.S. National Library of Medicine, 31 July 2023, www.ncbi.nlm.nih.gov/books/NBK518982/#:~:text=An%20 epidural%20 hematoma%20(EDH)%20 is,sutures)%20where%20the%20dark%20 inserts.

Khan, Sahil, et al. “Recent Advancements in Pathogenesis, Diagnostics and Treatment of Alzheimer’s Disease.” Current Neuropharmacology, U.S. National Library of Medicine, Nov. 2020, www.ncbi.nlm.nih.gov/pmc/articles/PMC7709159/.

Khellaf, Abdelhakim, et al. “Recent Advances in Traumatic Brain Injury.” Journal of Neurology, U.S. National Library of Medicine, Nov. 2019, www.ncbi.nlm.nih.gov/pmc/articles/PMC6803592/.

Knight, James, and Appaji Rayi. “Transtentorial Herniation.” StatPearls [Internet]., U.S. National Library of Medicine, 17 July 2023, www.ncbi.nlm.nih.gov/books/NBK560536/.

Knight, James, and Orlando De Jesus. “Tonsillar Herniation.” StatPearls [Internet]., U.S. National Library of Medicine, 23 Aug. 2023, www.ncbi.nlm.nih.gov/books/NBK562170/#:~:text=Tonsillar%20 herniation%20is%20the%20 movement,of%20 patients%20with%20his%20 condition.

Kostecki, Kassie, et al. “Subfalcine Herniation.” StatPearls [Internet]., U.S. National Library of Medicine, 23 Aug. 2023, www.ncbi.nlm.nih.gov/books/NBK536946/.

Kuns, Brianne, et al. “Memantine.” StatPearls [Internet]., U.S. National Library of Medicine, 31 Jan. 2024, www.ncbi.nlm.nih.gov/books/NBK500025/#:~:text=Memantine%20is%20an%20 antagonist%20 of,excitatory%20 neurotransmitter%20in%20the%20 brain.

La Russa, Raffaele, et al. “Post-Traumatic Meningitis Is a Diagnostic Challenging Time: A Systematic Review Focusing on Clinical and Pathological Features.” International Journal of Molecular Sciences, U.S. National Library of Medicine, 10 June 2020, www.ncbi.nlm.nih.gov/pmc/articles/PMC7312088/#:~:text=Post%2Dtraumatic%20meningitis%20refers%20to,trauma%2C%20regardless%20of%20temporal%20proximity.

Laplante, Jennifer, and Kristen St. George. “Antiviral Resistance in Influenza Viruses: Laboratory Testing.” Clinics in Laboratory Medicine, Elsevier, 12 Apr. 2014, www.sciencedirect.com/science/article/abs/pii/S0272271214000249.

Lazaridis, Christos, and Brandon Foreman. “Management Strategies Based on Multimodality Neuromonitoring in Severe Traumatic Brain Injury.” Neurotherapeutics, Elsevier, 25 July 2023, www.sciencedirect.com/science/article/pii/S1878747923019906.

Lemus, Hernan Nicholas, et al. “Multiple Sclerosis: Mechanisms of Disease and Strategies for Myelin and Axonal Repair.” Neurologic Clinics, Elsevier, 17 Nov. 2017, www.sciencedirect.com/science/article/pii/S0733861917300774?via%3Dihub.

Ling, Helen, et al. “Neurological Consequences of Traumatic Brain Injuries in Sports.” Molecular and Cellular Neuroscience, Academic Press, 12 Mar. 2015, www.sciencedirect.com/science/article/pii/S104474311500041X?via%3Dihub.

Ma, Heather M, and Ross D Zafonte. “Amantadine and Memantine: A Comprehensive Review for Acquired Brain Injury.” Taylor and Francis Online, Brain Injury, 20 Feb. 2020, www.tandfonline.com/doi/full/10.1080/02699052.2020.1723697.

Maas, Andrew, et al. “Traumatic Brain Injury: Progress and Challenges in Prevention, Clinical Care, and Research - the Lancet Neurology.” Traumatic Brain Injury: Progress and

Challenges in Prevention, Clinical Care, and Research, National Center for Biotechnology Information, 29 Sept. 2022, www.thelancet.com/journals/laneur/article/PIIS1474-4422(22)00309-X/fulltext.

Maiese, Kenneth. “Brain Herniation - Brain Herniation.” Merck Manual Professional Edition, Merck Manual Professional Edition, Apr. 2024, www.merckmanuals.com/professional/neurologic-disorders/coma-and-impaired-consciousness/brain-herniation.

Mao, Gordon. “Skull Fracture - Skull Fracture.” Merck Manual Consumer Version, Merck Manual Consumer Version, Mar. 2023, www.merckmanuals.com/home/injuries-and-poisoning/head-injuries/skull-fracture.

McGrath, Ailbhe, and Roger S Taylor. “Pediatric Skull Fractures.” StatPearls [Internet]., U.S. National Library of Medicine, 23 Jan. 2023, www.ncbi.nlm.nih.gov/books/NBK482218/.

Mckee, Ann C, and Daniel H Daneshvar. “The Neuropathology of Traumatic Brain Injury.” Handbook of Clinical Neurology, U.S. National Library of Medicine, 2015, www.ncbi.nlm.nih.gov/pmc/articles/PMC4694720/.

Mena, Jorge Humberto, et al. “Effect of the Modified Glasgow Coma Scale Score Criteria for Mild Traumatic Brain Injury on Mortality Prediction: Comparing Classic and Modified Glasgow Coma Scale Score Model Scores of 13.” The Journal of Trauma, U.S. National Library of Medicine, Nov. 2011, www.ncbi.nlm.nih.gov/pmc/articles/PMC3217203/.

Meredith, Gloria E, et al. “A Critical Review of the Development and Importance of Proteinaceous Aggregates in Animal Models of Parkinson’s Disease: New Insights into Lewy Body Formation.” Parkinsonism & Related Disorders, Elsevier, 17 Mar. 2004, www.sciencedirect.com/science/article/abs/pii/S1353802004000021.

Mesfin, Fassil B, et al. “Diffuse Axonal Injury.” StatPearls [Internet]., U.S. National Library of Medicine, 12 June 2023, www.ncbi.nlm.nih.gov/books/NBK448102/.

Miron, Veronique E, et al. “Oligodendrocyte.” Cells of the Oligodendroglial Lineage, Myelination, and Remyelination, ScienceDirect, Feb. 2011, www.sciencedirect.com/science/article/pii/S0925443910002103.

Mokri, Bahram. “The Monro-Kellie Hypothesis.” Neurology , American Academy of Neurology, 26 June 2001, www.neurology.org/doi/10.1212/WNL.56.12.1746#:~:text=What%20finally%20come%20to%20be,both%20of%20the%20 remaining%20two.

Mu, Jiao, et al. “Myelin Damage in Diffuse Axonal Injury.” Frontiers in Neuroscience, U.S. National Library of Medicine, 19 Mar. 2019, www.ncbi.nlm.nih.gov/pmc/articles/PMC6433984/#:~:text=It%20is%20understood%20traditionally%20that,vulnerability%20of%20oligodendrocytes%20in%20DAI.

Munakomi, Sunil, and Joe M Das. “Brain Herniation.” StatPearls [Internet]., U.S. National Library of Medicine, 13 Aug. 2023, www.ncbi.nlm.nih.gov/books/NBK542246/.

Muzio, Maria Rosaria. “Histology, Axon.” StatPearls [Internet]., U.S. National Library of Medicine, 14 Nov. 2022, www.ncbi.nlm.nih.gov/books/NBK554388/.

"NEUROCOGNITIVE OUTCOMES." Gulf War and Health Volume 7: Long-Term Consequences of Traumatic Brain Injury Institute of Medicine, Board on Population Health and Public Health Practice, Committee on Gulf War and Health: Brain Injury in Veterans and Long-Term Health Outcomes, vol. 7, National Academies Press, Washington, D.C, 2009.

Office of the Commissioner. "Traumatic Brain Injury: What to Know." U.S. Food and Drug Administration, FDA, 23 Aug. 2021, www.fda.gov/consumers/consumer-updates/traumatic-brain-injury-what-know-about-symptoms-diagnosis-and-treatment.

Patsopoulos, Nikolaos A. "Genetics of Multiple Sclerosis: An Overview and New Directions." Cold Spring Harbor Perspectives in Medicine, U.S. National Library of Medicine, 2 July 2018, www.ncbi.nlm.nih.gov/pmc/articles/PMC6027932/.

Pavel, Jaroslav, et al. "Neural Tissue Loss after Spinal Cord Injury." Cellular, Molecular, Physiological, and Behavioral Aspects of Spinal Cord Injury: The Neuroscience of Spinal Cord Injury, Academic Press, pp. 187–197. ScienceDirect, https://www.sciencedirect.com/science/article/abs/pii/B9780128224274000162. Accessed 27 July 2024.

Pengo, Marta, and Ferdinando Squitieri. "Beyond CAG Repeats: The Multifaceted Role of Genetics in Huntington Disease." Genes, U.S. National Library of Medicine, 19 June 2024, www.ncbi.nlm.nih.gov/pmc/articles/PMC11203031/.

Rajashekar, Devika. "Intracerebral Hemorrhage." StatPearls [Internet]., U.S. National Library of Medicine, 6 Feb. 2023, www.ncbi.nlm.nih.gov/books/NBK553103/#:~:text=Introduction,morbidity%20and%20mortality%5B1%5D.

Sanchack, Kristian E, and Craig A Thomas. "Autism Spectrum Disorder: Primary Care Principles." American Family Physician, American Family Physician, 15 Dec. 2016, www.aafp.org/pubs/afp/issues/2016/1215/p972.html.

Santos, Sara Figueiredo, et al. "Post-Traumatic Meningitis in Children: Eleven Years' Analysis." Acta Medica Portuguesa, U.S. National Library of Medicine, 12 Aug. 2011, pubmed.ncbi.nlm.nih.gov/22015025/.

Sharma, Rohit. "Subdural Hemorrhage: Radiology Reference Article." Radiopaedia, Radiopaedia.org, 11 Feb. 2024, radiopaedia.org/articles/subdural-haemorrhage?lang=us.

Shenoy, Saraswati Satyanarayan, and For Shing Lui. "Neuroanatomy, Ventricular System." StatPearls [Internet]., U.S. National Library of Medicine, 24 July 2023, www.ncbi.nlm.nih.gov/books/NBK532932/#:~:text=The%20ventricular%20system%20of%20the,times%2C%20its%20 function%20was%20obscure.

Skjeldal, Ola H, et al. Long-Term Post-Concussion Symptoms, Tidsskriftet, 22 Aug. 2022, tidsskriftet.no/en/2022/08/clinical-review/long-term-post-concussion-symptoms.

Sonne, James, et al. “Neuroanatomy, Substantia Nigra.” StatPearls [Internet]., U.S. National Library of Medicine, 24 Oct. 2022, www.ncbi.nlm.nih.gov/books/NBK536995/#:~:text=The%20 substantia%20 nigra%20(SN)%20is,Cognitive%2C%20and%20 Limbic%20 Outputs).

Su, Erik, et al. “Diffuse Axonal Injury.” Translational Research in Traumatic Brain Injury, CRC Press, Boca Raton, Florida, 2016, https://www.ncbi.nlm.nih.gov/books/NBK326722/. Accessed 30 July 2024.

Tenny, Steven, and William Thorell. “Intracranial Hemorrhage.” StatPearls [Internet]., U.S. National Library of Medicine, 17 Feb. 2024, www.ncbi.nlm.nih.gov/books/NBK470242/.

Trafton, Anne. “Seeing the Brain’s Electrical Activity.” MIT News | Massachusetts Institute of Technology, Massachusetts Institute of Technology, 26 Feb. 2018, news.mit.edu/2018/seeing-brains-electrical-activity-0226.

Traub, Stephen J, and Eelco F Wijicks. “Intraparenchymal Hemorrhage.” Intraparenchymal Hemorrhage - an Overview | ScienceDirect Topics, Emergency Medicine Clinics of North America, 2016, www.sciencedirect.com/topics/medicine-and-dentistry/intraparenchymal-hemorrhage.

“Types of Brain Injuries: Traumatic vs. Non-Traumatic.” Edited by Kenneth Ngo, Brooks Rehabilitation, Brooks Rehabilitation, 9 Apr. 2024, brooksrehab.org/conditions/brain-injury/types/.

Wang, Kevin K, et al. “An Update on Diagnostic and Prognostic Biomarkers for Traumatic Brain Injury.” HHS Public Access, NCBI, Feb. 2018, www.ncbi.nlm.nih.gov/pmc/articles/PMC6359936/pdf/nihms-1514604.pdf.

Wang, Ling, et al. “Autism Spectrum Disorder: Neurodevelopmental Risk Factors, Biological Mechanism, and Precision Therapy.” International Journal of Molecular Sciences, U.S. National Library of Medicine, 17 Jan. 2023, www.ncbi.nlm.nih.gov/pmc/articles/PMC9915249/.

WHO. “Meningitis.” World Health Organization, World Health Organization, 17 Apr. 2023, www.who.int/news-room/fact-sheets/detail/meningitis#:~:text=Meningitis%20is%20the%20inflammation%20of,transmitted%20from%20person%20to%20person.

Xiong, Ye, et al. “Emerging Treatments for Traumatic Brain Injury.” Expert Opinion on Emerging Drugs, U.S. National Library of Medicine, Mar. 2009, www.ncbi.nlm.nih.gov/pmc/articles/PMC2773142/#:~:text=Recent%20reviews%20 have%20 identified%20 several,%2C%20and%20 rivastigmine%20%5B25%5D.

Zafar, Saman, and Sridhara S Yaddanapudi. “Parkinson Disease.” StatPearls [Internet]., U.S. National Library of Medicine, 7 Aug. 2023, www.ncbi.nlm.nih.gov/books/NBK470193/.

Øverland, Elisabeth, et al. “Exploring Life with Autism: Quality of Life, Daily Functioning and Compensatory Strategies from Childhood to Emerging Adulthood: A Qualitative Study Protocol.” Frontiers, Frontiers, 14 Nov. 2022, www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2022.1058601/full.

Discoveries and Challenges in South Asian Ancient DNA Research: Unveiling Migration Patterns and Ancestral Admixture By Aarav Patel

Abstract

There are emerging studies into the admixture and interactions of the Ancestral North Indians and the Ancestral South Indians within South Asia. Ancient DNA (aDNA) is DNA gathered from remains dated to a long time ago, consequently aDNA can be used to help clarify historical questions, like the interactions and admixture of the Ancestral North Indians and the Ancestral South Indians within South Asia. This paper aims to answer how studies of aDNA have helped to clarify the migration patterns of Ancestral North and South Indians into and within South Asia. Previous research has shown that most populations in South Asia today descend from two major groups, the Ancestral North Indians and the Ancestral South Indians. In this paper, a multitude of different studies and articles are examined in order to compile major findings in South Asian genetic history, especially in regard with Ancestral North and South Indians. Some of the findings in this paper include the genetic substructure of the Indian Cline, the effect of endogamy and the caste system on the genetic landscape of South Asian populations, and the challenges of doing aDNA research in South Asia. These findings shed light into South Asia's history, the ancestry of several South Asian populations, and how aDNA research can be used to provide ancient perspectives on the genetic topography of South Asia.

Introduction

For decades, the origins and migration patterns of ancient South Asian populations have been shrouded in obscurity. Now, advancements in genetic research, particularly the study of ancient DNA (aDNA), are shedding light on how ancestral North and South Indian groups migrated, settled, and interacted across the South Asian landscape. The region of South Asia has remarkable genetic diversity, with some groups possessing African ancestry, others implying a degree of European ancestry, and other groups still resembling Southeast Asians genetically (Thangaraj and Rai). In addition, a great number of populations are unique to the Indian Subcontinent (Thangaraj and Rai). There exists a number of questions about how ancestral South Asian populations made their way onto the subcontinent. Some of these questions include: Where did Ancestral North Indians and Ancestral South Indians come from; How did the Harappan Civilization related genetically to later South Asian populations; The impacts of intermarrying only within castes (called endogamy) on the genetic variation of South Asia; etc. (Thangaraj and Rai).

Two major groups populated South Asia: the Ancestral North Indians and the Ancestral South Indians (Reich et al.). The Ancestral North Indians (ANI) are genetically related to populations residing in Central Asia, Southwest Asia, and Europe (Reich et al.). In fact, they formed the basis of the Indo-European languages spoken in the subcontinent (Reich et al.). In contrast, the Ancestral South Indians (ASI), based on current research, seem to be indigenous to the subcontinent and are genetically distinct, with them forming the basis of the Dravidian

languages spoken in the subcontinent (Reich et al.). Differences in language appear to align with ANI and ASI ancestry patterns, with Indo-European speakers having a higher percentage of ANI ancestry, and Dravidian speakers having a higher percentage of ASI ancestry (Reich et al.). Furthermore, upper-caste groups tend to have higher proportions of ANI ancestry compared to lower or middle castes (Reich et al.). Initially, when the ANI moved into South Asia, they began intermarrying with the ASI, leading to the formation of the Indian Cline, a gradient of populations which range from increasing ANI heritage to increasing ASI heritage (Reich et al.). The Indian Cline represents the non-uniform admixture (when unrelated populations mix together, creating a population with a blended ancestry) of the ANI and ASI genomes influenced by factors such as geography, cultural practices, and social structures (Reich et al.). The settlement of South Asia by the ANI and ASI led to distinct genetic and cultural patterns, with linguistic and social divisions reflecting their ancestry, while their admixture formed the Indian Cline.

A tool that can be used to reconstruct the migration history of those two groups is aDNA. ADNA is DNA that is gathered from sources such as ancient bones, teeth, and sediment (Sarkissian et al.). Working with aDNA can be often challenging as samples are often degraded, fragmented, or contaminated by modern DNA (Sarkissian et al.). Geneticists can then use aDNA to create a better picture on populations that migrated and mixed in South Asia as it can provide a pure genome of certain ancestral groups, which can help create a better picture on how the admixture between the ANI and ASI occurred. The process in which the genome represented within aDNA is read is called genetic sequencing (Sarkissian et al.). ADNA can be sequenced in a variety of ways, such as shotgun sequencing or single-stranded DNA library preparation (Sarkissian et al.). Shotgun sequencing is a method used to sequence aDNA in which the DNA is randomly broken into small fragments which are then read separately (Sarkissian et al.). Then, the resulting sequences are reassembled using computational tools to reconstruct the entire genome (Sarkissian et al.). In contrast, single-stranded DNA library preparation involves converting DNA into a single stranded form, typically by denaturing it (Sarkissian et al.). Using this method, researchers can efficiently sequence shorter DNA fragments, therefore increasing the chances of successfully sequencing aDNA (Sarkissian et al.). ADNA is being collected from ancient human remains found at archeological sites in South Asia to shed light on past migration and admixture, especially between the ANI and ASI populations. The aim of this paper is to investigate how genetic studies of aDNA have clarified the migration patterns of Ancestral North and South Indians into and within South Asia.

Genetic studies that helped to clarify migration of ANI and ASI in and to South Asia

One of the major groups that populated South Asia were the Ancestral North Indians (ANI), who spoke Indo-European languages and were not indigenous to South Asia. One tool that researchers use to determine the ancestry of the ANI is y-chromosomal analysis (Thanseem et al.; Reich et al.). Y-chromosomal analysis is the sequencing and analysis of the y-chromosome in males, which can be used to trace back paternal lineages much farther than autosomal

sequencing, as the y-chromosome is not subject to crossing over during meiosis, leading to a more stable chromosome that changes less overtime and that can be traced back to one ancestor (Thanseem et al.; Reich et al.). One major result from y-chromosome analyses of South Asian populations is that ANI ancestry is more prevalent in the y-chromosome (Thanseem et al.; Reich et al.).

When comparing autosomal estimates of ANI ancestry to genetic characteristics of West Eurasians, there is a correlation on the y-chromosome (Reich et al.). In contrast, there is a marginal correlation using mitochondrial DNA (mtDNA), implying that there was a stronger male flow of the ANI people into South Asia (Reich et al.). Mitochondrial DNA analysis can be used by researchers to trace back maternal lineages much farther back than autosomal sequencing, as the mitochondria in a person's cells are ultimately derived from the mother's egg cell (Thanseem et al.; Thangaraj and Rai). Additionally, because mitochondrial DNA changes very slowly over time, it can be traced back to a group of ancestors (Thanseem et al.). Mitochondrial DNA changes slowly over time as it is not subject to crossing over during meiosis like autosomal chromosomes, thus allowing for a genetic sequence that represents ancient mitochondrial DNA (Thanseem et al.). A significant portion of the ancestry of ANI may have been derived from the Eurasian steppe—a vast region of unforested grassland (steppe) stretching from modern-day Hungary to Manchuria—as modern-day groups that have significant ANI ancestry, such as the Ror people also have a significant proportion of genetic similarities to the peoples residing in the Eurasian Steppe (Pathak et al.). In addition, steppe ancestry seems to be one of the main ancestral components of the Indian Cline—a gradient between mostly ANI and mostly ASI relatedness—with most people on the Indian Cline also having some proportion of steppe ancestry (Kerdoncuff et al.; Moorjani et al.; Reich et al.). Some populations in South Asia also have an additional Middle-Eastern component in their genome, reflecting a Middle-Eastern, ANI, and ASI ancestry (Kumar et al.). This further highlights that additional groups might have had an admixture with the Indian Cline after they settled in South Asia, increasing the genetic complexity of that area. Overall, these studies show that the ANI had ancestry in the Eurasian Steppe, with that ancestry still being noticeable among South Asian populations today.

The second major group already settled in South Asia during the ANI and ASI admixture event were, of course, the Ancestral South Indians (ASI). They spoke Dravidian languages and were indigenous to India during the ANI migration into South Asia. In the modern-day, South Asian populations with the most ASI ancestry continue to speak Dravidian languages, contrasting with the Indo-European languages spoken by populations with greater ANI ancestry. The ASI population are not related to any groups outside India, seeming to have arisen from a relatively small population with long periods of isolation and limited gene flow (Tamang and Thangaraj). However, due to the aforementioned admixture event between the ANI and ASI, which spread throughout the whole of South Asia, the modern populations of South Asia with the most ASI ancestry still have some proportion of ANI ancestry (Tamang and Thangaraj). Among the modern ASI-related groups, most of them speak Dravidian languages, which seem to be indigenous to South Asia (ArunKumar et al.; Tamang and Thangaraj). In contrast, the

Indo-European languages spoken by the ANI, share a common linguistic ancestry with languages spoken in West Asia and Europe (ArunKumar et al.; Tamang and Thangaraj). One tool which has been used to determine the ancestry of the ASI is mitochondrial DNA analysis. One major trend found by using mitochondrial DNA analysis on South Asian populations is that maternal lineages found on the Indian Cline more commonly link back to ASI populations than their y-chromosome analysis counterpart (Tamang and Thangaraj; Thanseem et al.; Reich et al.). This suggests that the ANI migration into India was a mostly male-mediated migration that created modern South Asian populations with a paternal ANI ancestry and a maternal ASI ancestry (Thanseem et al.; Tamang and Thangaraj). Overall, these studies show that the ASI had ancestry that is indigenous to South Asia, with that ancestry still being noticeable among South Asian populations today.

The admixture of the ANI and ASI populations led to the formation of the Indian Cline reflecting the fact that different South Asian populations have inherited differing proportions of ancestry from both the ANI and ASI (Moorjani et al.; Reich et al.). This admixture could have occurred around 1,856 to 4,176 years ago, which highlights that most major South Asian populations experienced some measure of admixture within the last few thousand years, including more isolated groups (Moorjani et al.). One reason why the Indian Cline has persisted for so long is due to social stratification caused by the caste system and endogamy. As the Indian caste system promotes endogamy (marrying within caste), each caste can be classified as its own distinct subgroup with a unique proportion of ANI and ASI ancestry (Sengupta et al.; Bamshad et al.). Furthermore, the practice of endogamy has served to fossilize the genomes of caste subgroups and the wider Indian Cline, giving a picture of an admixture of two populations in progress (Sengupta et al.; Bamshad et al.). Each caste subgroup is genetically distinct, with subgroups within a wider group often having more genetic variance than those wider groups with each other (Sengupta et al.). However, there are some groups that are completely off the Indian Cline. One of these are the Andaman islanders, who appear to have completely no ANI ancestry, representing a picture on how an original ASI population could look (Tamang and Thangaraj). Additionally, there are groups that speak Austroasiatic languages that trace back their ancestry to a population that was not the ANI or the ASI. There is evidence that suggests that the original Austroasiatic speaking population originally came from East or Southeast Asia (Tamang and Thangaraj; Tätte et al.). Overall, these studies show that there is a gradient of genetic ancestry between the ANI and ASI in modern South Asian populations. All the studies in this section have cleared up and shed more light on the migration of the ANI into South Asia and the subsequent admixture event between the ANI and ASI.

Genetic studies investigating admixture and population structure of the ANI and ASI populations

The ANI-ASI admixture event was a complex process involving a major migration event of the ANI people into South Asia, and the consequent interactions that occurred. Some populations, such as the Pandit population in the Jammu and Kashmir region of South Asia,

exhibit a high degree of ANI ancestry, indicating that the ANI population migrated from the north of South Asia (Reich et al.). Other populations, such as the Chenchu people in Andhra Pradesh, have a high degree of ASI ancestry, suggesting that the ASI people had the highest concentration in the southern portion of South Asia after the ANI migration event (Reich et al.). Lastly, some groups in the Andaman and Nicobar Islands exhibit a pure ASI-adjacent ancestry with no ANI ancestry at all, suggesting that the ASI population reached as far as the Andaman and Nicobar Islands, but the ANI population didn't reach as far (Reich et al.). Overall, these studies show that different populations with different geographical locations have differing proportions of ANI and ASI ancestry—with more northern populations having more ANI ancestry, and more southern populations having more ASI ancestry.

One reason that the Indian Cline has remained distinct today is because of the prevalence of endogamy perpetuated by the caste system. Endogamy is a marriage practice in which one marries someone of their own social group—caste in this instance (Sengupta et al.; Reich et al.; Bamshad et al.). Endogamy in South Asia stratified the population, locking in the genetic gradient of the Indian Cline and preventing the ANI and ASI admixture event to progress further (Sengupta et al.; Reich et al.; Bamshad et al.). One trend among caste subgroups within a larger population is that the higher caste subgroups tend to have an increased proportion of ANI ancestry, and lower caste subgroups tend to have a greater proportion of ASI ancestry (Bamshad et al.). This implies that during the admixture event that led to the formation of the Indian Cline, the ANI people set themselves at the top of the social hierarchy—with position based on proportion of ANI ancestry—leading to the modern-day caste system in South Asia (Sengupta et al.; Reich et al.; Bamshad et al.). This is significant because it posits that the caste system served as a way to differentiate those with higher proportions of ANI ancestry from those with lower proportions of ANI ancestry, thus keeping subgroups with heavier ANI ancestry higher on the social hierarchy of South Asia.

One method of contrast between caste subgroups is high melanin (skin tone) diversity (Nizamuddin et al.; Sengupta et al.). Traditionally, a fairer skin tone has been used to distinguish high caste populations, and vice versa for lower caste populations (Nizamuddin et al.; Sengupta et al.). This may have come about due to ANI people having a fairer skin tone than the ASI population, thus when population structures stratified due to the proliferation of endogamy, the higher caste subgroups found themselves with fairer skin tones—reminiscent of the ANI people—and lower caste subgroups having a darker skin tone—reminiscent of the ASI people (Nizamuddin et al.; Sengupta et al.). This implies that skin tone has been used as a marker to further differentiate different caste groups from one another, thus increasing the stratification of South Asian populations. The studies in this section have helped to clarify on how the admixture of the ANI and ASI populations took place.

Challenges and limitations in studying the aDNA of South Asian populations

ADNA is a noteworthy source of data for studying the admixture of the ANI and ASI populations. However, there are some challenges in obtaining viable aDNA samples, such as

environmental conditions. In particular, the warm climate of South Asia can hamper the ability to gain viable aDNA samples, as the warm temperatures increases the damage to DNA after death even further than usual due to the heat (Thangaraj and Rai; Dabney et al.). Additionally, the warmer temperatures increase the probability that microorganisms and intracellular nucleases (enzymes within a cell that degrade DNA and RNA) will damage DNA and cause fragmentation and reduction in the overall DNA sample (Dabney et al.). This underscores how aDNA studies and databases based on South Asian populations may be limited.

Currently, there is not a comprehensive database of aDNA from South Asia (Chakraborty and Basu; Thangaraj and Rai). Some reasons for why aDNA databases with South Asian samples are limited is because of the relative scarcity of aDNA samples from South Asia, and a focus on European aDNA finds (Thangaraj and Rai; Dabney et al.). Of the South Asian aDNA studies that exist, there are substantial limitations in their data. One of those studies is a sequencing of aDNA from bone fragments under a St. Augustine church in Goa that allegedly belonged to Queen Ketavan of Georgia (Thangaraj and Rai). The results of that study proved the speculation that those bone fragments were of Georgian origin (Thangaraj and Rai). Another study involved the sequencing of aDNA from the skeletal remains of Parsi refugees in the site of Sanjan (Thangaraj and Rai). Lastly, a study was conducted on aDNA samples of Harappan civilization remains from archeological sites such as Rakhigarhi, Haryana, and Dholavira, Gujarat (Thangaraj and Rai). Those aDNA studies are few and far between, with isolated aDNA samples that are only marginally connected with the greater Indian Cline. This highlights the need for more comprehensive aDNA databases from South Asia, as it can help in uncovering the genetic history of South Asian populations.

Another challenge in successfully studying aDNA from South Asia is the contamination of older archeological sites. As aDNA is fragmented and can contain DNA from other sources, such as microorganisms, after the aDNA is sequenced, it is filtered by comparing it to a closely related sequenced reference genome (Peyrégne and Prüfer). However, similar exogenous DNA sequences that are also related to the reference genome can pass through this computational filtering, thus causing contamination of aDNA datasets (Peyrégne and Prüfer). There have been many archeological sites in South Asia that have been discovered to may have aDNA that have been contaminated due to previous excavation and research on finds found at those given sites that did not utilize proper procedures such as wearing protective equipment, being careful with remains and artifacts, and placing remains in sealed containers (Thangaraj and Rai; Chakraborty and Basu; Peyrégne and Prüfer).

Lastly, aDNA samples should be obtained in an ethical manner that pays importance to the historical and cultural relevance of given archeological finds. Many different cultures emphasize the importance of honoring the dead and thus not disturbing human remains. Given this context, in October of 2021, a group of researchers, archeologists, curators, and geneticists from 24 different countries published *Ethics of DNA Research on Human Remains: Five Globally Applicable Guidelines* which highlights five guidelines any prospective aDNA researcher should use when procuring samples (Alpaslan-Roodenberg et al.). First, that

researchers should follow all rules and regulations from where they are procuring samples from. Second, researchers should create a detailed plan of action before any study on human remains. Third, researchers should minimize damage to human remains. Fourth, researchers should make all data found from human remains available following publication. However, one should also take care to not release data openly if it is in conflict with Indigenous Data Sovereignty (Kowal et al.; Alpaslan-Roodenberg et al.). And fifth, that researchers should communicate with stakeholders (groups or individuals that have a connection with the archeological finds, i.e. descendent groups or people responsible for stewardship of those human remains) from the beginning of the study and to ensure respect and sensitivity for those stakeholder's perspectives. This paper was a landmark publication concerning the ethical issues of aDNA research on human remains, which is a major part of this paper. The studies in this section have helped to highlight on the current challenges and limitations of conducting aDNA research in South Asia.

Conclusion

ADNA has helped to clarify the origins, admixture, and population dynamics that took place during the various migration events in the region of South Asia (Thangaraj and Rai; Reich et al.). The two major ancestral components of the present-day Indian Cline comprised of ANI and ASI, respectively, standing for Ancestral North Indians and Ancestral South Indians, which make up much of the present-day genetic variability within South Asians (Thangaraj and Rai; Reich et al.). The ANI population was closely related to the peoples of Central Asia and the Eurasian Steppe and has contributed substantially to the paternal ancestry of South Asia (Reich et al.; Thangaraj and Rai; Thanseem et al.). ASI, however, can be interpreted to represent the indigenous people of the subcontinent, with more connections with ancient hunter-gatherers, and the origin of most of the maternal ancestry of present-day South Asians (Thangaraj and Rai; Tamang and Thangaraj; Reich et al.; Thanseem et al.).

The two, ANI and ASI together, form a genetic continuum called the Indian Cline, wherein a gradient formed by extensive admixture is seen (Reich et al.). Over time, geographical and social structures such as endogamy and the caste system halted further gene flow, freezing the genetic substructure in the Indian population (Bamshad et al.; Reich et al.). This allowed researchers an unparalleled view of how population admixtures affect the genetic topography of the affected populations, and specifically allowed researchers to further extrapolate the ancestry and genetic uniqueness of the ANI and ASI populations.

The hot climate and microbial contamination of South Asia render recovering aDNA quite problematic (Thangaraj and Rai; Dabney et al.). This has reduced the quantity and quality of aDNA in South Asian genetic databases, which deprives researchers of valuable data that can be useful for further research on South Asian genomes. Setting this aside, technological advances have made it possible for researchers to retrieve aDNA from human remains dating several thousand years and give insight into the ANI and ASI admixture event (Thangaraj and Rai; Reich et al.). However, there are related ethical issues surrounding the practice of using human remains to gather aDNA (Alpaslan-Roodenberg et al.; Kowal et al.). Nevertheless, studies of

aDNA continue to add depth to the complex genetic history of South Asia by studying how ancient migrations, environmental factors, and cultural practices may influence our knowledge of historical migrations into South Asia.

The findings discussed in this paper have served to address several questions on the historical and cultural understanding of South Asia. First, they shed light on the ancestry of the ANI and ASI peoples. They also clarify on how the ANI and ASI admixture event created the Indian Cline, a gradient of ANI to ASI ancestry across the subcontinent. However, they have posed new questions as well, such as how external populations, like West Asian and Austroasiatic populations contributed to the genetic landscape of South Asia, and what steps are necessary to build a more comprehensive database of South Asian aDNA. Additionally, another major question was how researchers could balance scientific goals with ethical considerations of respecting the culture and history of different peoples. These findings have disproven many assumptions about South Asian ancestry, such as the idea that the Harappan civilization was completely disconnected from later South Asian populations, that modern South Asian ancestry is rigidly divided into distinct ASI and ANI groups, and the "Aryan Invasion Theory" which set forth the idea that a large-scale, violent invasion by Indo-European-speaking people was the primary source of ANI ancestry. Overall, genetic studies of aDNA have served to clarify ANI and ASI migration patterns and provide insight on the finer details of the ANI and ASI admixture event.

Building on these findings, technology and sequencing techniques of aDNA have advanced far in recent years, further enhancing the ability to trace ancestry, reconstruct migration patterns, and deepen our understanding of admixture events and social structures across South Asia (Sarkissian et al.; Reich et al.). Practices such as shotgun sequencing and single-stranded library DNA preparation helped to properly sequence short and degraded aDNA fragments (Sarkissian et al.). Also, y-chromosomal and mitochondrial DNA analysis offers the ability to trace ancestry with greater clarity (Reich et al.; Thanseem et al.). Lastly, new computational tools help to filter out contaminated DNA (Sarkissian et al.; Dabney et al.). These advanced technologies could help to reconstruct migration patterns, shed more light on admixture events, give insights into social structures and when they formed, and to highlight connections between broader populations.

Genetic insights given by aDNA have helped to shift our perspective on South Asian population movements and admixture, with it clarifying the ANI and ASI admixture, helping to shed light on how population stratification has affected South Asian genomes, and illustrating South Asian peoples genetic connection to broader populations. This is essential since understanding ANI-ASI admixture, population stratification, and broader genetic connections helps to reconstruct the complex history of South Asian populations, provides insights into their cultural, historical, and genetic development, and challenges outdated conceptions about their ancestry.

Works Cited

Alpaslan-Roodenberg, Songül, et al. "Ethics of DNA Research on Human Remains: Five Globally Applicable Guidelines." *Nature*, vol. 599, no. 7883, Nov. 2021, pp. 41–46. *PubMed Central*, https://doi.org/10.1038/s41586-021-04008-x.

ArunKumar, GaneshPrasad, et al. "Population Differentiation of Southern Indian Male Lineages Correlates with Agricultural Expansions Predating the Caste System." *PLoS ONE*, vol. 7, no. 11, Nov. 2012, p. e50269. *pmc.ncbi.nlm.nih.gov*, https://doi.org/10.1371/journal.pone.0050269.

Bamshad, Michael, et al. "Genetic Evidence on the Origins of Indian Caste Populations." *Genome Research*, vol. 11, no. 6, June 2001, p. 994. *pmc.ncbi.nlm.nih.gov*, https://doi.org/10.1101/gr.173301.

Chakraborty, Saikat, and Analabha Basu. "Reconstruction of Ancestral Footfalls in South Asia Using Genomic Data." *Journal of Biosciences*, vol. 44, no. 3, July 2019, p. 74. *DOI.org (Crossref)*, https://doi.org/10.1007/s12038-019-9875-5.

Dabney, Jesse, et al. "Ancient DNA Damage." *Cold Spring Harbor Perspectives in Biology*, vol. 5, no. 7, July 2013, p. a012567. *cshperspectives.cshlp.org*, https://doi.org/10.1101/cshperspect.a012567.

Kerdoncuff, Elise, et al. "50,000 Years of Evolutionary History of India: Insights from ~2,700 Whole Genome Sequences." *bioRxiv*, Feb. 2024, p. 2024.02.15.580575. *pmc.ncbi.nlm.nih.gov*, https://doi.org/10.1101/2024.02.15.580575.

Kowal, Emma, et al. "Community Partnerships Are Fundamental to Ethical Ancient DNA Research." *Human Genetics and Genomics Advances*, vol. 4, no. 2, Apr. 2023, p. 100161. *ScienceDirect*, https://doi.org/10.1016/j.xhgg.2022.100161.

Kumar, Lomous, et al. "Genetic Affinities and Adaptation of the South-West Coast Populations of India." *Genome Biology and Evolution*, vol. 15, no. 12, Dec. 2023, p. evad225. *pmc.ncbi.nlm.nih.gov*, https://doi.org/10.1093/gbe/evad225.

Moorjani, Priya, et al. "Genetic Evidence for Recent Population Mixture in India." *American Journal of Human Genetics*, vol. 93, no. 3, Sept. 2013, p. 422. *pmc.ncbi.nlm.nih.gov*, https://doi.org/10.1016/j.ajhg.2013.07.006.

Nizamuddin, Sheikh, et al. "Skin Pigmentation Diversity in Central East Indian Populations and Its Correlation with Mitochondrial Haplogroups." *Genetics & Genomic Sciences*, vol. 1, no. 1, Dec. 2016, pp. 1–7. *DOI.org (Crossref)*, https://doi.org/10.24966/GGS-2485/100005.

Pathak, Ajai K., et al. "The Genetic Ancestry of Modern Indus Valley Populations from Northwest India." *American Journal of Human Genetics*, vol. 103, no. 6, Dec. 2018, p. 918. *pmc.ncbi.nlm.nih.gov*, https://doi.org/10.1016/j.ajhg.2018.10.022.

Peyrégne, Stéphane, and Kay Prüfer. "Present-Day DNA Contamination in Ancient DNA Datasets." *BioEssays*, vol. 42, no. 9, 2020, p. 2000081. *Wiley Online Library*, https://doi.org/10.1002/bies.202000081.

Reich, David, et al. "Reconstructing Indian Population History." *Nature*, vol. 461, no. 7263,

Sept. 2009, p. 489. *pmc.ncbi.nlm.nih.gov*, https://doi.org/10.1038/nature08365.
Sarkissian, Clio Der, et al. "Ancient Genomics." *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 370, no. 1660, Jan. 2015, p. 20130387. *pmc.ncbi.nlm.nih.gov*, https://doi.org/10.1098/rstb.2013.0387.
Sengupta, Dhriti, et al. "Population Stratification and Underrepresentation of Indian Subcontinent Genetic Diversity in the 1000 Genomes Project Dataset." *Genome Biology and Evolution*, vol. 8, no. 11, Nov. 2016, p. 3460. *pmc.ncbi.nlm.nih.gov*, https://doi.org/10.1093/gbe/evw244.
Tamang, Rakesh, and Kumarasamy Thangaraj. "Genomic View on the Peopling of India." *Investigative Genetics*, vol. 3, no. 1, Oct. 2012, p. 20. *BioMed Central*, https://doi.org/10.1186/2041-2223-3-20.
Tätte, Kai, et al. "The Genetic Legacy of Continental Scale Admixture in Indian Austroasiatic Speakers." *Scientific Reports*, vol. 9, no. 1, Mar. 2019, p. 3818. *www.nature.com*, https://doi.org/10.1038/s41598-019-40399-8.
Thangaraj, K., and Niraj Rai. "Peopling of India: Ancient DNA Perspectives." *Journal of Biosciences*, vol. 44, no. 3, July 2019, p. 70. *DOI.org (Crossref)*, https://doi.org/10.1007/s12038-019-9874-6.
Thanseem, Ismail, et al. "Genetic Affinities among the Lower Castes and Tribal Groups of India: Inference from Y Chromosome and Mitochondrial DNA." *BMC Genetics*, vol. 7, no. 1, Aug. 2006, p. 42. *BioMed Central*, https://doi.org/10.1186/1471-2156-7-42.

Improving Resource Management with Precision Agriculture: Water and Soil Analysis on a Temecula Avocado Farm By Vikram Anand

Abstract

Due to a surge of climate-related disasters in the last decade, California farmers are experiencing increased pressure to make ends meet. Precision agriculture, also referred to as smart farming or precision farming, is the practice of resource application on a need-basis to enhance yield. This practice allows farmers to increase production while cutting costs at the same time. Despite these benefits, farmers are hesitant to implement smart farming practices due to upfront costs and a lack of knowledge about potential benefits. After installing an automated smart farming system on a seven-acre avocado farm in Temecula, California, this study gathered water usage data every month over a year-long period. Next, an in-depth analysis of the data measured the economic and resource-usage benefits of precision agriculture. An irrigation audit along with a soil test was conducted through a local extension office to measure the farm's current water usage efficiency, pressure levels, and determine ways to increase yield and reduce waste. The results of the study indicated that an automated watering system is around 37.79% less costly on a monthly basis than a manual system and uses 45.08% less water, or 62.22 less HCF (hundred cubic feet) of water monthly. Comparing the original cost and the savings, it would take just 23.65 months to recover the cost of the system. Precision agriculture has the potential to significantly lower utility costs and water usage. These smart farming practices can be applied to farms across California experiencing similar drought issues.

Key Terms

Precision Agriculture, Irrigation, Water Management, Sustainable Farming

Introduction

In the context of a rapidly changing climate, Southern California farmers are experiencing extreme weather events like droughts and floods, which may impact their crop production. To safeguard against this unpredictability, farmers must use their resources as efficiently as possible. This is where precision agriculture comes into play. Precision agriculture is a way for farmers to focus special attention on their crops. By using modern technologies like sensors or drones, farmers can improve their crop productivity and resource management systems. However, due to a combination of knowledge and monetary barriers, many farmers are unable to benefit from these practices. To measure the benefits of precision agriculture, an experiment was conducted to demonstrate the scalable profitability offered by smart farming practices. This paper examines how precision agriculture and efficient resource usage can assist small-scale farms in growing high-value crops, specifically avocados.

Overview of Precision Agriculture

Smart farming refers to the advanced management of agriculture, where farmers provide resources and nutrients to a given site on a farm based on what it needs and how it responds to abiotic factors, such as water, temperature, and sunlight. Precision agriculture also encourages site-specific treatment to optimize resource management (Mandal and Ghosh). The process can increase yield, as plants receive needed resources rather than a generalized application of fertilizers and water. Smart farming can further improve crop yield by allowing farmers to use different data readings, such as soil moisture, water retention, water usage, and more, to increase a farm's profit margins. Some common methods of precision agriculture used today include mechanized irrigation and site-specific nutrient application.

History of Usage

Precision agriculture is not a new practice, it has simply changed with the advancement of technology. Many ancient farms emphasized the intense management of crops on a small plot to generate maximum yield (Johnson et al.). More recently, the 20th century saw a push towards generalized farming, or low-intensity care on a large plot of land. This meant applying the same treatment and nutrients over a large area to minimize labor costs per field. The main goal of this practice was to maximize efficiency while still producing a decent crop yield (Lowenberg-DeBoer). However, generalized farming was a flawed system, as the increased efficiency sacrificed optimal crop production in the process. The contemporary rise of precision agriculture, pushed by research and environmentally-focused organizations, once again encourages site-specific treatment. More and more farmers are switching to precision agriculture because of its cost benefits. Today, advanced technology is being incorporated into precision agriculture. Complex datasets are created to generate a mechanized path forward for a farm (Johnson et al.). Devices such as GPS, sensors, drones, and more, are used commercially to manage resources and data. A complete adoption of precision agriculture would see a fully mechanized system where manual labor is cut down as much as possible (Levy). Coinciding with this, some drawbacks of a mechanized farm include a reduced need for actual farmers, which could result in a loss of farming jobs in rural areas.

Benefits

One of the major benefits of precision agriculture is that it offers farmers a return on their investment in the long term. These cost savings, combined with the resources it saves make it well worth it once implemented (Levy). Some other common goals of smart farming include increased yield, pest or disease management, and improved monitoring systems. The benefits of using smart farming are significant, as farmers can save both money and resources while still increasing their yield. Smart farming also has the potential to decrease resource usage and environmental waste in a big way if many farmers start to implement it. Precision agriculture has further benefits, such as eliminating excess labor and resource costs while saving on farm expenses. This modernized way of farming is still a relatively new idea to many farmers. One of the biggest issues in the spread of smart farming is the lack of knowledge that many farmers

have of smart farming (Walter et al.). This combined with the initial price can make it look like a risky investment to many farmers. That is the reason why so far, precision agriculture is largely only implemented in large, industrial farms (Walter et al.). However, precision agriculture is becoming increasingly more affordable as the cost of mechanization goes down (Gopikrishna). So far, specific nutrient and water applications have mostly been used on high-yield, low-cost crops such as corn and wheat, to minimize risk.

Methodology

Automated System

The experiment took place on a farm in Temecula, California, and occurred over a year-long period for data comparison. Prior to the installation of the automated system on the avocado farm, the crops were irrigated on average once a week in 24-hour cycles. The frequency of this practice changed based on an assessment of the weather and rainfall in the area. This method of water usage caused excessive waste. To fix this issue, an automated irrigation system was implemented on the farm. To do this, a smart irrigation company, IRRIOT, was contacted to purchase the parts shown in Figure 1 that would create the system shown in Figure 2.

Item	Price	Quantity	Cost
Controller	$2022.33	1	$2022.33
Wifi Dongle	$251.42	1	$251.42
RTU-2	$480.99	3	$1442.96
Installation Kit	$32.79	3	$98.38
Annual subscription	$262.36	1	$262.36
Ceramic high amplifier antenna with cable	$174.90	1	$174.90
Volumetric Soil Moisture and Temp Sensor BGT-SM1	$185.84	2	$371.67
			Total Cost: $4624.02

Figure 1: Parts and cost of the precision agriculture system

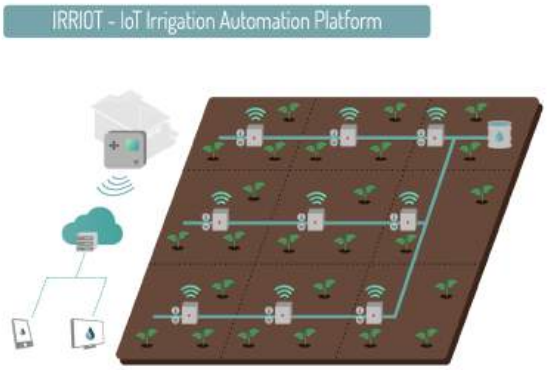


Figure 2: Flow of communication through the IRRIOT system

The main parts of the system are the Controller, RTU (Remote Terminal Unit), and the soil moisture sensors. The controller is the base machine that controls all of the valves and sprinklers. The system works by the RTU communicating with the controller to turn specific watering lines on and off based on the soil moisture data that the sensor collects. The sensor is constantly taking data, which in turn automatically waters the trees on a twice-a-week schedule whenever the soil around the trees dips below a certain soil moisture percentage. Through the implementation of precision agriculture onto the farm, the clear positive effect it had on water usage became evident when compared from both a monetary and water usage standpoint to the previous year. The total cost of the original automation system is shown above in Figure 1. The complete system was implemented on August 24, 2024, as shown in Figure 3.

Figure 3: Full system and control hub installed onto the farm

Soil Test

The next step for the farm is to improve resource usage efficiency further. One way that this is possible is by improving the farm's soil health. While efficient water usage is a very important part of managing resources, ensuring the soil is healthy and retains the water is just as

important. As the soil holds more water and is healthier, less water is used to properly irrigate the crops. To determine the next steps for the grove and how the increase in soil health can influence resource management, the TEAM RCD conservation office visited the farm and took a soil sample from two feet deep near the avocado tree to determine the overall soil composition and the nutrients it possesses. This included nitrogen levels, phosphorus levels, potassium levels, and pH levels.

Irrigation Audit

Another part of improving resource use efficiency is to maximize the uniformity of watering systems. To do this, the TEAM RCD conservation office once again came out to the farm and carried out an irrigation audit to assess the functionality of the farm's management systems. The audit report measured pressure uniformity and flow uniformity to offer a complete analysis of the grove's current watering efficiency along with steps that could be taken to improve the overall resource management.

Results

Results of the Automated System

The comparison of water usage from 2023 to 2024 after the implementation of the IRRIOT automated irrigation system is shown below in Figure 5.

Month	Water Usage HCF(Gallons)	Water Cost
September. 2023 (No System)	166 HCF (124,177 gallons)	$603.92
October 2023 (No System)	110 HCF (82,285 gallons)	$431.10
Sept. 2024 (System Installed)	79 HCF (59,096 gallons)	$381.00
October 2024 (System Installed)	72.57 HCF (54286.13 gallons)	$262.97

Figure 5: Water usage 2023 vs. 2024

As is displayed in the table, there was a $222.92 decrease in water usage from September 2023 to September 2024 and a $168.13 decrease from October 2023 to October 2024. Based on this, the average monthly water cost decreased by about 37.79%. Overall, water usage decreased by around 45.08% when averaged across both months over the two years.

Return On Investment

Using the average monthly savings of the two tested months from Figure 5 ($195.53), the data was extrapolated and used to calculate the irrigation system's return on investment over

different time periods. The breakdown of how much money the system saves over different periods compared to the original cost of the system is shown below in Figure 6.

	1 Month Cost/Water Savings ($/HCF)	2 Month Cost/Water Savings ($/HCF)	1-Year Cost/Water Savings ($/HCF)	2-Year Cost/Water Savings ($/HCF)	5 Years Cost/Water Savings ($/HCF)	10 Years Cost/Water Savings ($/HCF)
	$195.53 (62.22 HCF)	$391.06 (125 HCF)	$2,346.36 (747 HCF)	$4,692.72 (1,493.28 HCF)	$11,731.80 (3,735 HCF)	$23,463.60 (7,470 HCF)
System Cost	$4624.02	$4624.02	$4624.02	$4624.02	$4624.02	$4624.02
Total ROI on System (Savings - System Cost)	-$4,428.49	-$4,037.43	-$1,691.07	$68.7	$7,107.78	$16,355.82

Figure 6: Return on investment over different time periods

Through the calculations, the study found that it would take about 23.65 months, or around two years, to break even on the initial investment, saving thousands of dollars over a five or ten-year period. Not only does purchasing this system make economic sense, but it also highlights how much water can be saved on even a relatively small farm in Southern California through the use of precision agriculture.

Results of the Soil Test

The report from the soil tests conducted by TEAM RCD reported the nutrient levels that exist in the soil near the avocado trees, as shown in Figure 7.

Nutrient/Metric	**Level**	**Supplementation Recommendations**
Nitrate nitrogen (N)	Low	3x a year - Spring, summer, and fall with fertilizer.
Phosphorus (P)	Low	Use slow-release fertilizer in a regulated fertilizing schedule
Potassium (K)	Low	Supplemental potassium can be applied on a regular fertilizing schedule
Conductivity (soil salts)	Normal (Lower values are generally better)	Control pH and salt levels
pH	Normal	The application of gypsum and sulfur will increase yield

Figure 7: Results of soil test from the conservation office

Due to the low levels of Nitrogen, Phosphorus, and Potassium in the soil, the conservation office recommended that the best course of action to take would be to apply a slow-release 16-16-16 or 15-15-15 N-P-K (Nitrogen-Phosphorus-Potassium) fertilizer to the field, improving crop yield along with soil fertility.

Results of the Irrigation Audit

The results of the irrigation audit conducted by the TEAM RCD conservation office measured both pressure uniformity and flow rate, which were then used to determine emission uniformity.

Pressure uniformity measures how stable the pressure output and water uniformity at each sprinkler valve is. This indicates which sprinklers use too much water due to high pressure and which avocado trees are not getting enough water due to low pressure. Generally, a pressure of around 30 psi is considered good. The results of the test are displayed below in Figure 8.

Highest Inlet Pressure	Lowest Inlet Pressure	Average System Pressure	Overall Pressure Uniformity
50 psi	35 psi	43 psi	81%

Figure 8: Pressure statistics measured by audit

Flow uniformity measures how much water comes out of the sprinklers per hour. This is an important metric because it directly calculates the water usage of each main line. The closer the farm's highest and lowest flow rates are, the better the flow rate uniformity will be. The results of the audit are displayed below in Figure 9.

Highest Flow Rate	Lowest Flow Rate	Average System Flow	Overall Flow Uniformity
32.4 gallons per hour	2.5 gallons per hour	13.9 gallons per hour	66%

Figure 9: Flow rate statistics measured by audit

The emission uniformity is a combined measurement of both the flow uniformity and the pressure uniformity. This offers a holistic value that calculates overall resource usage efficiency. Based on both the pressure uniformity and flow uniformity, the farm's emission uniformity lies at 73%. Based on industry standards, the uniformity ranks in the "fair" range. While this value is decent, a farm with good resource usage and management would have an emission uniformity in the 90-100% range. This shows that the total farm system performance, measured in emission uniformity, is greatly impacted by the various problems that appear in the farm's watering system. The report stated the problems shown in Figure 10 after the assessment.

Problem #1	Problem #2	Problem #3
The low flow rate at the top of the hill is caused by a large loss of pressure	Pressure regulators are missing throughout the water lines.	Mixed sprinkler types appear throughout the grove.

Figure 10: Problems stated by the audit report

Based on the aforementioned problems, the audit offered a few steps forward, shown in Figure 11, that would allow the farm to save resources and achieve a much higher emission uniformity rating.

Recommendation #1	Recommendation #2	Recommendation #3
The farm needs new valve-specific pressure-compensating micro-sprinklers	Install more sensors at different depths to offer more data on which to make decisions.	Sprinklers should be of the same type on the farm and should be 18 gallons per hour.

Figure 11: Recommendations offered by the audit report

Discussion

As the trials demonstrated, precision agriculture offers promising results in saving water and minimizing costs. The automated system significantly improved the farm's water usage and the implementation of certain nutrient fertilizers will help boost overall crop growth. Another benefit of the system was improved monitoring systems throughout the farm. The results of this experiment illustrate that when done correctly, precision agriculture can greatly reduce farmers' water costs. For farmers suffering from similar issues due to the impacts of drought, the introduction of mechanized systems and site-specific nutrient processes could potentially improve yield and save money on crops and water. Furthermore, increasing water efficiency and reducing water waste is a big step in adapting to the impacts of climate change. Precision agriculture could be the key to a more sustainable future of agriculture, especially if supported by state and federal policies. Increased legislative support, in the form of incentives and cost-sharing programs, would allow for the mass implementation of advanced technologies and automated watering systems onto farms.

Conclusion

Precision agriculture has the potential to act as a tool for efficient resource management in today's agriculture. The practices are easy to maintain and learn about, along with being environmentally friendly and economically beneficial to farmers who implement these technologies on their farms. Using an in-depth experiment involving the implementation of an automated water control system on a 7-acre California Farm, I found that water usage decreased by nearly 45%, $195.53, or 62.22 HCF on average. Through these savings, the system will pay

for itself in just 23.65 months. This not only displays the scalable monthly savings but also the relatively short time it takes to break even, despite the initial large investment. In addition, the use of an irrigation audit and soil test further emphasized the need and benefits of site-specific resource management. Conducting this experiment on the farm illustrated a few aspects of precision agriculture and how it can change the landscape of future agriculture. The full use of smart farming would include practices such as cover cropping and manure application to maximize yield and reduce resource waste on a global level. Precision agriculture is not limited to this farm, but rather it extends all across California to farms that are experiencing droughts and excessive resource use at a time when it is imperative to save water. The introduction of automation in farming can change the environment and traditional farming practices for the better. Despite skeptics unsure about the feasibility of precision agriculture, smart farming has shown considerable growth in recent years due to its cost savings and positive environmental impacts. Smart agriculture is providing a way for farmers to save both money and resources in their practices. As the number of people who know about and implement these practices increases, precision agriculture could end up being the default farming system on both a small and commercial scale, changing the future of farms and food for the better.

Works Cited

Gopikrishna, S.R. “Soil Health and Support Systems: Contradictions and Missing Links.” Economic and Political Weekly, vol. 47, no. 29, 21 July 2012, pp. 24–26 www.jstor.org/stable/41720008.

Johnson, Richard, et al. “Precision Farming in Mechanized Agriculture.” SAE International, vol. 106, 1997, pp. 316–322, www.jstor.org/stable/44722762..

Levy, Wendy. “PRECISION AGRICULTURE: A Smart Farming Approach.” Spore, no. 185, 2017, pp. 4–7, www.jstor.org/stable/44242663.

Lowenberg-DeBoer, Jess. “The Precision Agriculture Revolution: Making the Modern Farmer.” Foreign Affairs, vol. 94, no. 3, 2015, pp. 105–112, www.jstor.org/stable/24483669.

Mandal, Debashis, and S. K. Ghosh. “Precision Farming – the Emerging Concept of Agriculture for Today and Tomorrow.” Current Science, vol. 79, no. 12, 2000, pp. 1644–1647, www.jstor.org/stable/24104120.

“Products.” Irriot, www.irriot.com/products/.

Walter, Achim, et al. “Smart Farming Is Key to Developing Sustainable Agriculture.” Proceedings of the National Academy of Sciences of the United States of America, vol. 114, no. 24, 2017, pp. 6148–6150, www.jstor.org/stable/26484181. 789

A DFT Analysis for Synthesizing Vitamin A By Tianyou Huang

Abstract

Vitamin A deficiency (VAD) is a major nutritional concern in lower-income countries. It is responsible for thousands of deaths in those countries every year. Thus, finding the optimal route for vitamin A synthesis is essential, especially for the countries that are influenced by VAD. Three mechanisms of synthesizing Vitamin A have been evaluated by Density Functional Theory (DFT) calculations. This experiment investigated the BASF $C_{15} + C_5$ Wittig approach, the Rhône-Poulenc $C_{15} + C_5$ Julia approach, and the Kuraray $C_{10} + C_{10}$ approach. The electronic energy, highest occupied molecular orbital energy, and dipole moments were calculated using the B3LYP functional and the 3-21g basis set. The energy profiles of these synthesis routes were compared to determine the most energetically favourable method. The Julia approach has the lowest energy change, indicating its higher efficiency in terms of energy compared to the Wittig and Kuraray methods. It is shown that other factors such as scalability and raw material availability should also be considered in industrial applications.

Keywords

Synthesis, Vitamin A, Density functional theory, Computational chemistry

Introduction

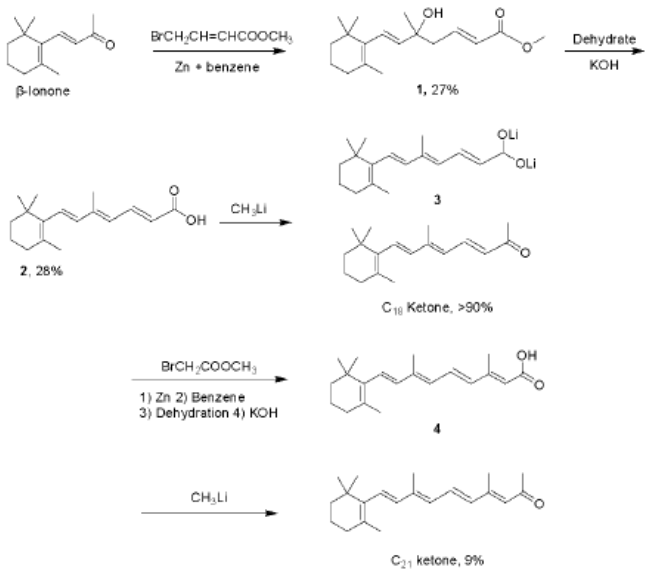
The proper operation of numerous critical metabolic and physiological processes, such as vision, immune system function, gene transcription, and skin cell differentiation, depends on vitamin A (Chapman, 2012). However, some diseases exist due to lack of vitamin A. Vitamin A deficiency (VAD) is a major nutritional concern in poor societies, especially in lower-income countries. Although the percentage is decreasing, in 2013, 30% of children under 5 years old were vitamin A deficient. About 2% of all deaths in this age group are attributable to this disease. Furthermore, vitamin A deficiency was found to be responsible for 94,500 deaths from diarrhoea and 11,200 deaths from measles in 2013. This accounted for 1.7% of all deaths in children under the age of five in low-income and middle-income countries (Stevens et. al., 2015). With such severe circumstances, a cost reduction in producing vitamin A would dramatically assist more people suffering from vitamin A deficiency.

The exploration of vitamin A structure and production has persisted for over 140 years. The finding of vitamin A might originate from a study conducted in 1816, the physiologist François Magendie experimented on dogs under nutritional deprivation, a technique that led to corneal ulcers and a high death rate. These results were comparable to the clinical condition commonly observed in undernourished, abandoned infants in Paris. Later, Frederick Gowland Hopkins postulated the existence of "unsuspected dietetic factors" as prerequisites for life in 1906 (Semba, 2012). In 1913, at the University of Wisconsin, Elmer Verner McCollum (1879–1967) and Marguerite Davis recorded an observation involving rats given "the ether extract of egg or of butter." They were given credit for finding what they dubbed "fat-soluble A,"

the first food additive to be identified as a vitamin (Rosenfeld, 1997). In 1920, Jack Cecil Drummond (1891–1952) suggested that because the substances are not all amines, they should be named Vitamin A, B, C, etc (Drummond, 1920). Vitamin A's structures were verified by Karrer, P., Morf, R., and Schöpp, K. later in 1931 (Karrer et. al., 1931). In 1937, Holmes and Corbet separated it and crystallized it from fish liver oils at Oberlin College (Holmes & Corbet, 1937).

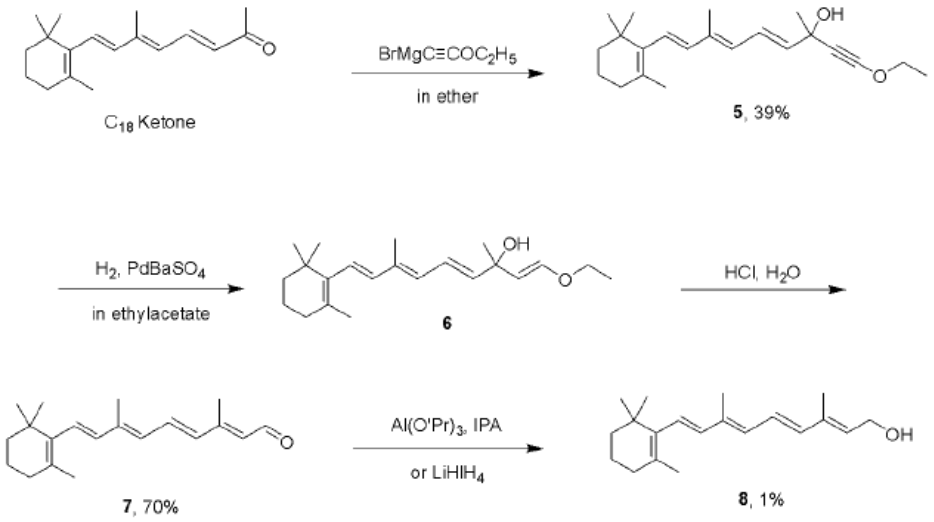
Van Dorp and Arens' synthesis of vitamin A acid and aldehyde

Retinoic acid (vitamin A acid) and retinol (vitamin A aldehyde) were first fully industrially synthesized in 1946 and 1947 by David Adriaan van Dorp (1915–1995) and Jozef Ferdinand Arens (1914–2001) (Arens & Dorp, 1946; Van-Dorp & Arens, 1947). The synthesis provided by Van Drop And Arens is a multistep synthesis from β-Ionone and undergoes the Reformatsky reaction, which symbolizes the process of forming β-hydroxy-esters by adding zinc enolates to aldehydes or ketones (Parker et. al., 2016). Three crystalline acids were isolated after the synthesis. The reaction mechanism that produces vitamin A acid is provided below (Scheme 1).



Scheme 1. Van Drop and Arens' Synthesis of Compounds with Strong Vitamin A Activity

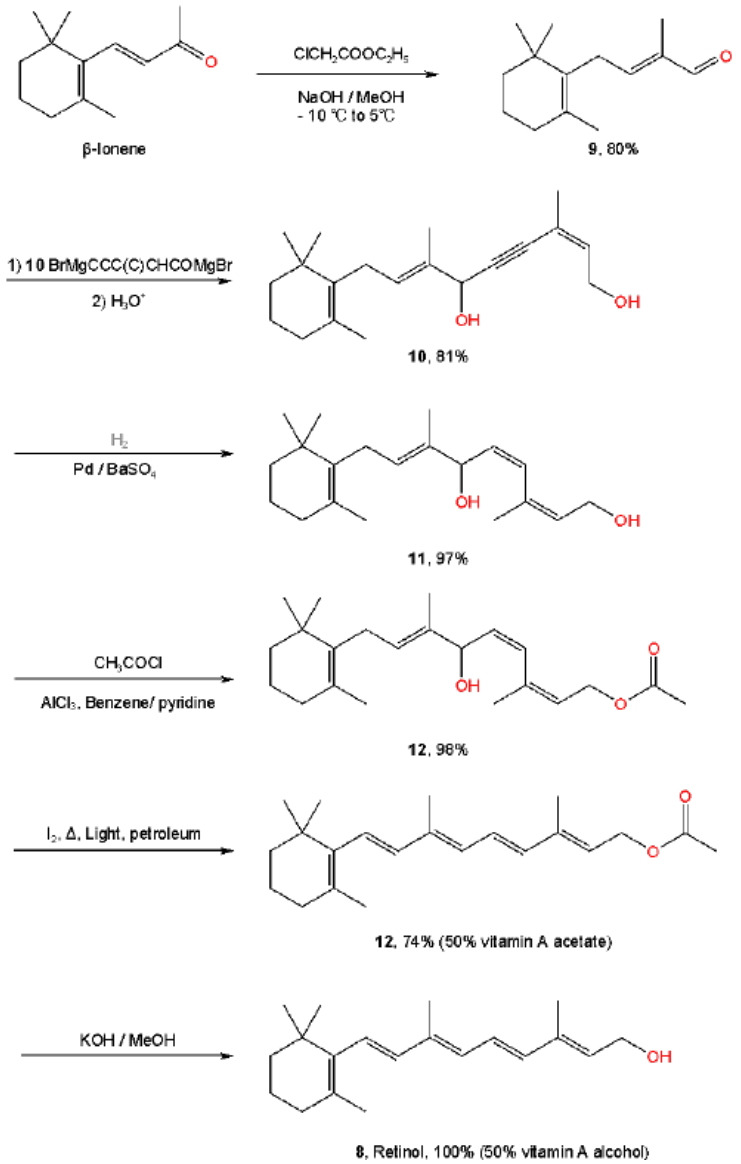
In 1947, vitamin A was synthesized by Van Dorp and Arens starting with the C_{18} Ketone that was synthesized in their 1946 study. The mechanism of the synthesization is provided below (Scheme 2). Vitamin A could be derived by reducing **7** using aluminium isopropoxide and isopropyl alcohol in a Meerwein-Ponndorf-Verley reduction, which 35% of the product is vitamin A. Later in 1949, Van Drop and Arens found that lithium aluminium hydride could also be used to do the reduction (Arens & Van-Dorp, 2010). This reaction path increased the production of vitamin A aldehyde to 50%. However, since it is time-consuming, it is not suitable for large-scale production (Parker et. al., 2016).



Schema 2. Arens-van Dorp reaction (Van-Dorp & Arens, 1947; Parker et. al., 2016)

O. Isler, W. Huber, A. Roneo und M. Kofler's synthesis of vitamin A

The reactions, similar to Van Dorp and Arens' synthesis, start with β-Ionone. However, since it avoids preparing ethoxy acetylene, it is not as time-consuming as the previous one and is thus suitable for large-scale production (Parker et. al., 2016). The reaction mechanism is shown in Scheme 3.

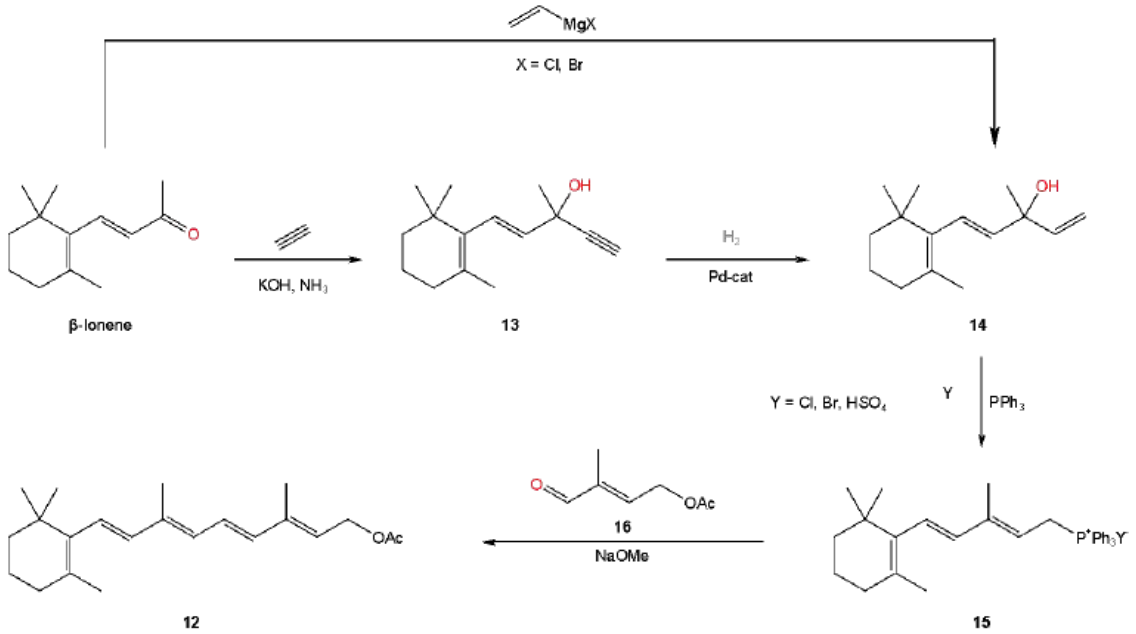


Scheme 3. O. Isler, W. Huber, A. Romeo und M. Kofler's Synthesis of Retinol (Parker et. al., 2016; Isler et. al., 1947)

Synthesizing vitamin A by Wittig Reaction

The reaction between benzophenone and methylenetriphenylphosphorane, subsequently known as the Wittig reaction or Wittig olefination, was discovered in the 1950s by Wittig and Geissler (Wittig & Geissler, 1953). Eventually, a method for synthesizing vitamin A was established, which relied on the Wittig olefination of a C_{15}-building block **15** with a C_5-moiety **16** (Sarnecki & Pommer, 2024). β-ionone was vinylated using a Grignard reaction (Arnould et. al., 1985). As an alternative, β-ionone can be semi hydrogenated and then added to acetylene to create vinyl-β-ionol **14** (Gould & Thompson, 1935). The corresponding C_{15}-phosphonium salt **15** is produced by treating triphenylphosphine with acid. Vitamin A acetate is produced by Wittig olefination of the second essential building block, a C_5-aldehyde **16**, with phosphonium salt

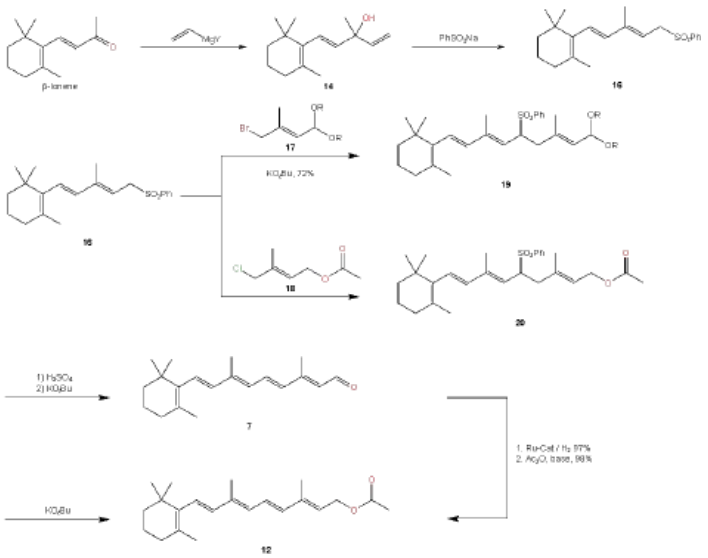
(Pommer & Nürrenbach, 1975). The coupling is carried out in aqueous solutions containing mild bases (such as ammonia or alkali metal carbonates) or in organic solvents like alcohols or DMF (Waddell & Chihara, 1981; Fisher et. al., 1974). To obtain the desired (all-E) vitamin A acetate, a mixture containing the (11Z)-isomer is obtained. This mixture can be photochemically induced to form the (all-E)-form in the presence of a photosensitizer and visible light, or it can be isomerized at elevated temperatures with Pd/C (Schleich & Stoller, 1978; Schaefer et. al., 2022; Mori & Nagashima, 1997). The reaction mechanism of synthesizing vitamin A acetate by Witting Reaction is shown in Scheme 4.



Scheme 4. Synthesis of Vitamin A acetate by Wittig reaction (Arnould et. al., 1985; Gould & Thompson, 1935)

Synthesis of vitamin A by Julia Chemistry

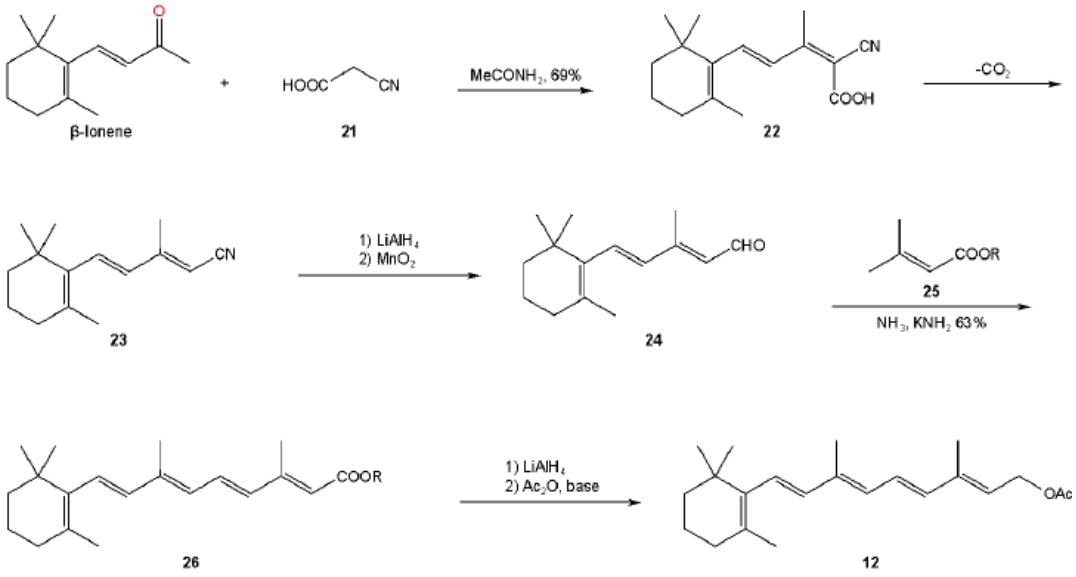
Rhône-Poulenc used the sulfone-based olefin synthesis, initially reported by Marc Julia in the 1970s, in an industrial process for the synthesis of vitamin A acetate (Julia, 1973). The reaction mechanism is shown in scheme 5. The C_{15}-sulfone 16 is produced by reacting β-Vinylol 14, which is derived from β-ionone, with sodium phenyl sulfonate (Julia, 1972). This is followed by a reaction with either bromo-acetal 17 or chloride 18 to produce the C_{20}-sulfones 19 and 20. Direct production of vitamin A acetate occurs from the removal of benzenesulfinic acid from 20. On the other hand, 19 is eliminated and hydrolysed to produce retinal 7, which is then reduced and acylated to produce vitamin A acetate (Décor, 1977; Chabardes et. al., 2024; Chabardes et. al., 1974; Grosselin et. al., 1991; Fischli & Mayer, 1975). The overall yield of the C_5-building block synthesis was assumed to be about 90% (Gould & Thompson, 1935).



Scheme 5. Synthesis of Vitamin A acetate by Marc Julia's theory

Synthesis of Vitamin A by Sumitomo

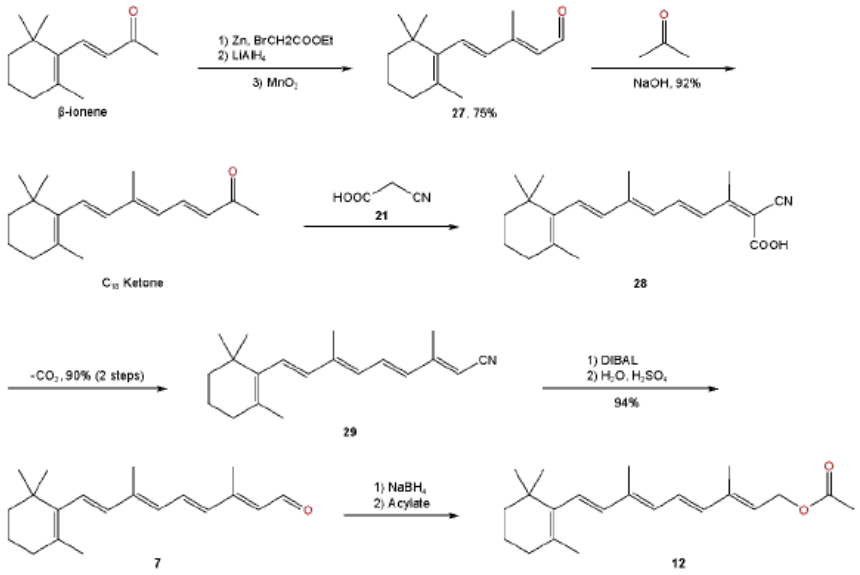
Sumitomo developed a synthesis pathway starting with β-ionone. In the reaction, vitamin A acetate is also formed by the coupling of C_{15} + C_5 building blocks. The mechanism is shown below (Scheme 6).



Scheme 6. Synthesis of Vitamin A Acetate by Sumitomo (Julia, 1972; Matsui et. al., 1958)

Synthesis of Vitamin A by Philips

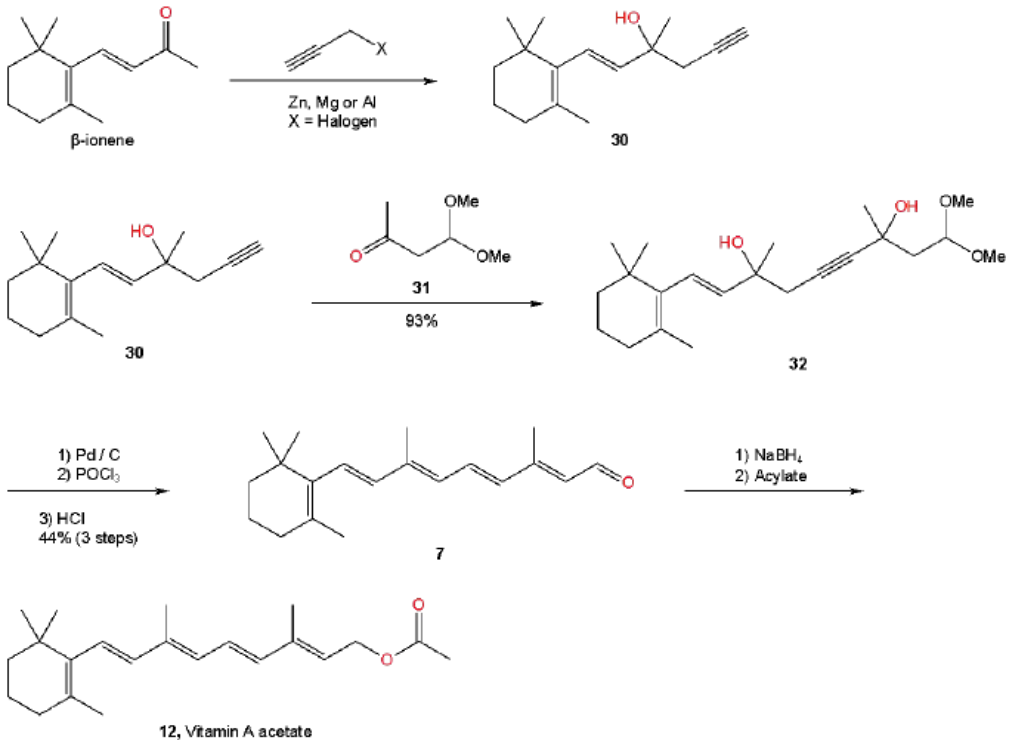
Philip's synthesis of vitamin A also starts with β-ionone and is the coupling of C_{15} + C_5 building blocks. However, its synthesis with C_{18}-ketone to form the corresponding C_{20}-nitrile **28** (Gould & Thompson, 1935). The mechanism is shown in scheme 7.



Scheme 7. Synthesis of Vitamin A acetate by Philips (Huisman & Smit, 1958; Huisman et. al., 2010)

Synthesis of Vitamin A by DPI and Glaxo

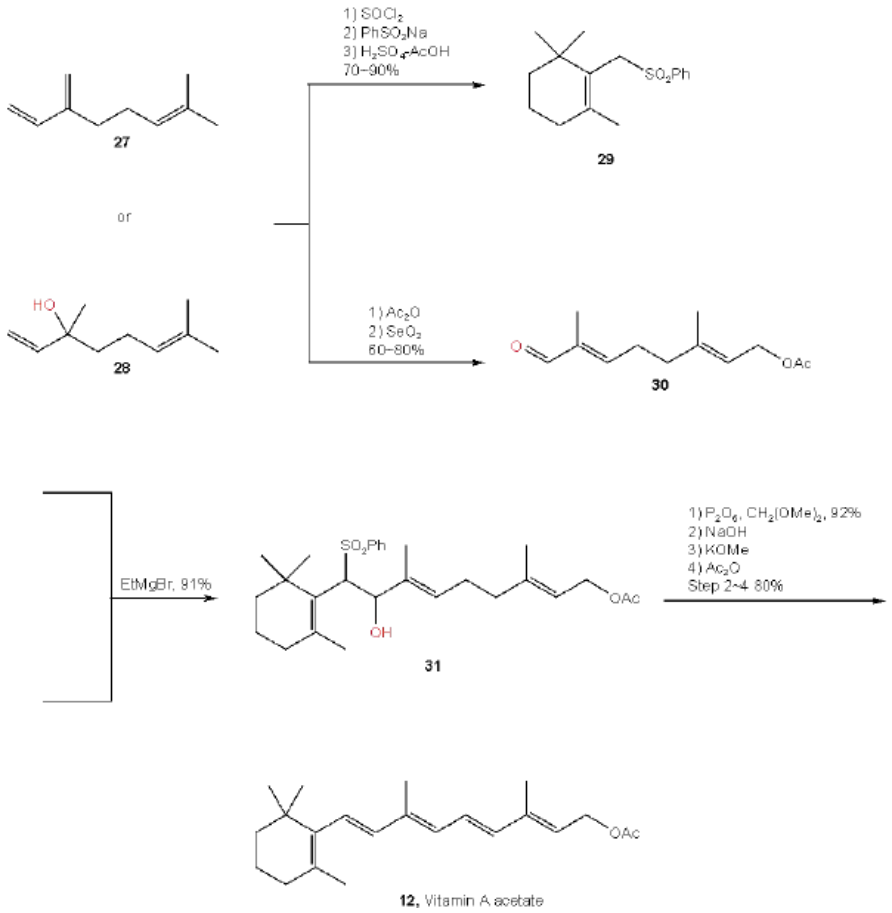
Distillation Products Limited and Glaxo in 1954, however, follows a $C_{16} + C_4$ approach. This synthesis has a reaction mechanism that is quite different from the previous ones that follow a $C_{15} + C_5$ approach (Scheme 8).



Scheme 8. Synthesis of Vitamin A Acetate by Distillation Products Limited and Glaxo (Humphlett, 1954; Humphlett & Burness, 1954)

Kuraray's synthesis of vitamin A

From the 1980s to 1990, the Japanese company Kuraray developed a pathway for synthesizing vitamin A by combining two C_{10} blocks. Two distinct C_{10}-compounds, sulfone **29** and aldehyde **30**, were produced from myrcene **27** or linalool **28**. When these two substances are combined, β-hydroxy-sulfone **31** is created, which is then transformed into vitamin A acetate. The mechanism is shown in Scheme 9.



Scheme 9. Synthesis of Vitamin A Acetate by Kuraray (Gould & Thompson, 1935; Lagerweij et. al., 1993; Otera et. al., 1989; Junzo et. al., 1986; Ladd, 1951; Mori et. al., 1991; Suzuki et. al., 1989; Otera et. al., 1988)

In conclusion, the experiment in this paper will mainly focus on three reaction mechanisms that produce vitamin A acetate: the BASF $C_{15} + C_5$ Wittig approach, Rhône-Poulenc/Adisseo $C_{15} + C_5$ Julia approach, and the Kuraray $C_{10} + C_{10}$ mechanism. The self-consistent field energy and highest occupied molecular orbital energy of intermediates in the mechanisms would be calculated by computational methods. An energy diagram of the production of Vitamin A through each mechanism will be drawn using the self-consistent field energy calculated and compared to find the optimal route from the three mechanisms.

Method

Density Functional Theory Fundamentals

Chemicals obey the laws of quantum mechanics, allowing predictions in-silico of chemical reactions (Townsend & Grayson, 2020). Quantum chemistry methods such as Density Functional Theory, aim to provide a solution to the Schrödinger equation, shown below:

$$\hat{H}\Psi(r_1, r_2,..., r_N) = E\Psi(r_1, r_2,..., r_N)$$

Where $\hat{H}$ is the Hamiltonian operator, E is the energy, Ψ is the wavefunction, and r_i is the coordinate of each electron (Jones, 2015). B By using the Kohn-Sham description of many-body systems, the electronic Schrödinger equation of chemical systems can be solved (Echenique & Alonso, 2007).

We have employed Density Functional Theory (DFT) is being used in this study. In physics and chemistry, DFT has been a major tool for investigating the electronic structure of periodic systems, like crystals, since the last 40 years (Jones, 2015). Potential energy surfaces (PES) of chemical systems can be computed with DFT.

DFT plays a crucial role in modeling chemical reactions by computing Potential Energy Surfaces (PES of chemical systems). These PES gives details about a chemical's energy at a variety of geometries and degrees of freedom (Jones, 2015). Analysis of PES enables the evaluation of transition states, activation barriers, or the energy differences between reactants and transition states. Thus, the DFT method described in this work has been used for calculating the electronic (Self-Consistent Field, SCF) energy of transition states to compare the practicability and efficiency of selected vitamin A synthetic routes.

Preparation

Structures in the calculations were prepared using RDKit from SMILES. The Python code to generate the structures can be found in Appendix 1.

All geometries and energies presented in this study were computed using the B3LYP functional theory (DFT) functional and the Grimme's D4 dispersion method as implemented in the ORCA Quantum Chemistry Package (version 6.0). B3LYP stands for "Becke, 3-parameter, Lee-Yang-Parr." This hybrid functional incorporates a portion of exact exchange energy from the Hartree-Fock theory along with exchange-correlation energies derived from other sources, specifically the local spin density approximation (LSDA) and the correlation functional developed by Lee, Yang, and Parr (LYP) (Lee et. al., 1988).

Geometry optimizations were performed using the 3-21G basis set, A split-valence double-ζ basis set. A double-ζ basis set consists of two basis functions for each atomic orbital, allowing for greater flexibility in representing the electron density around atoms. This contrasts with a minimal basis set, which uses only one function per orbital. The term "ζ" refers to the exponent in Slater-type orbitals (STOs), and having two functions enables better modelling of the variations in electron distribution due to chemical bonding and molecular interactions. The RIJCOSX method with auxiliary basis sets was used to speed-up the calculations. An example

of an optimized geometry employing this approach is shown in Figure 1.

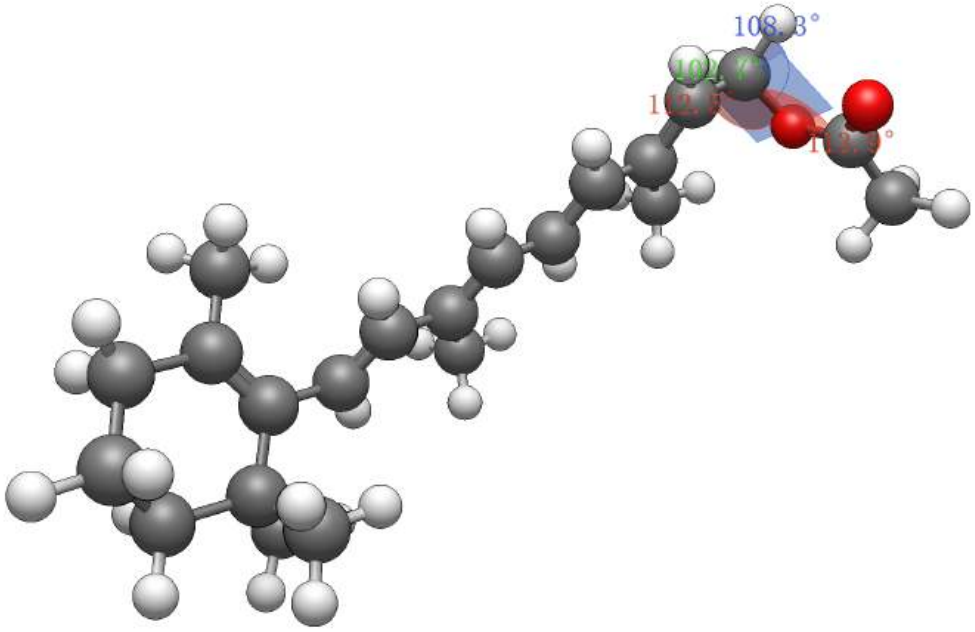


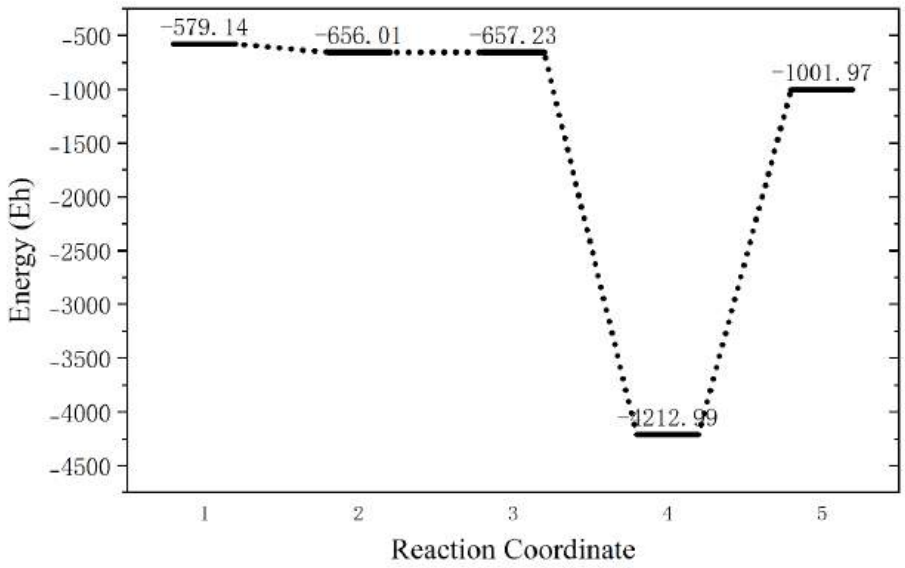
Figure 1. Optimized structure of vitamin A acetate

Self-Consistent Field (SCF) energies, orbital energies, and dipole moments were extracted from the ORCA output files (Neese, 2022; Neese, 2003; Neese & et. al., 2009; Bykov et. al., 2015; Stoychev et. al., 2017; Helmich-Paris et. al., 2021; Neese, 2023; Caldeweyher et. al., 2017; Caldeweyhe et. al., 2019; Caldeweyher et. al., 2020) .

Results

BASF C15 + C5 Wittig approach

The reaction mechanism of the DASF synthesis of vitamin A acetate by using the Wittig approach is already shown in Scheme 4. The following scheme is the reaction energy change of the vitamin A acetate synthesis by the Wittig reaction.



Scheme 10 Schematic representation (energy V.S. reaction coordinate) of the reaction mechanism by Wittig approach

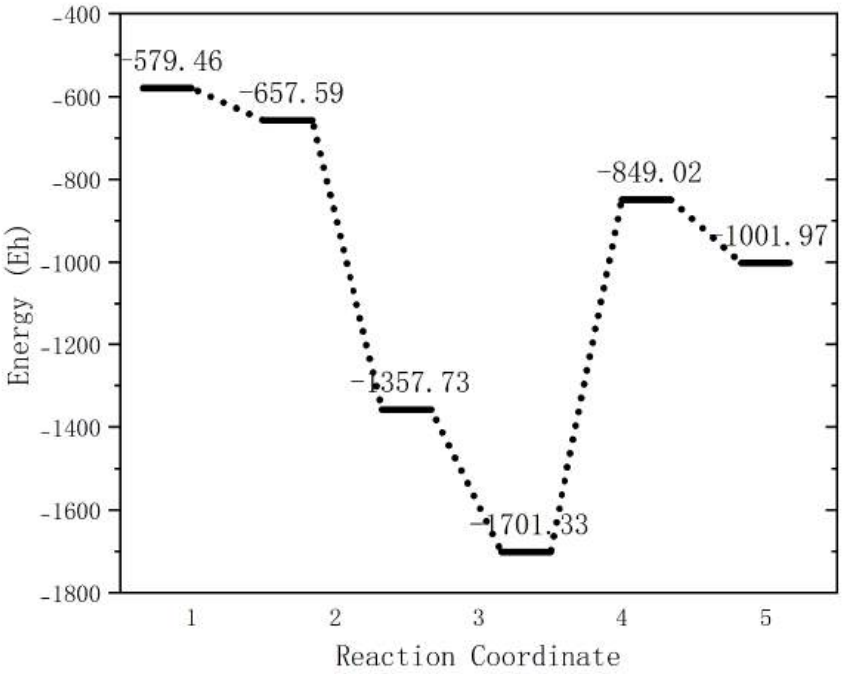
The specific final single energy, HOMO energy, and dipole magnitude are given in Table 1.

Table 1. Values calculated through the Density Functional Theory of synthesizing vitamin A by Wittig reaction

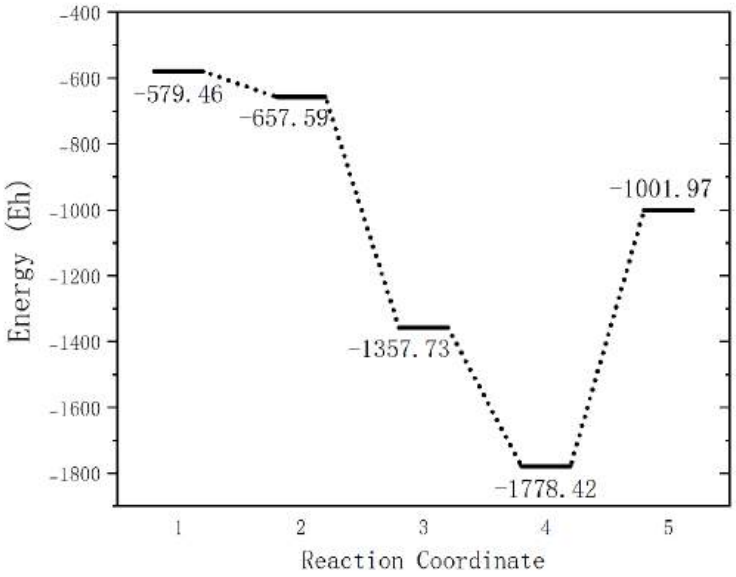
Compound name	***SCF Energy (Hartree)***	***HOMO energy (Hartree)***	***Dipole magnitude (Debye)***
β - Ionone	-579.13963	-0.218577	3.208074944
No.13	-656.01099	-0.208618	1.399691435
No.14	-657.22755	-0.195693	1.134328777
No.15	-4212.9864	-0.131678	12.07898961
Vitamin A Acetate	-1001.974713	-0.182186	1.794558566

Julia reaction (Rhône-Poulenc)

The reaction mechanism developed by Julia has two possible reaction routes, as shown in Scheme 5. The first path is through the direct synthesis of sulfone **20**. After eliminating the benzenesulfinic acid from compound **20** it will directly produce compound acetate **12**. The second path would include the synthesis of sulfone **19**. Sulfone **19** will go through both hydrolysis and elimination of benzenesulfinic acid to produce retinal **7**. Eventually, retinal **7** would convert to acetate **12** by acylation and reduction. The schematic representation of the energy of the routes is shown in schemes 11 and 12 separately. The specific final single energy, HOMO energy, and dipole magnitude are given in Tables 2 and 3.



Scheme 11. Schematic representation (energy V.S. reaction coordinate) of the first reaction mechanism developed by Marc Julia



Scheme 12. Schematic representation (energy V.S. reaction coordinate) of the second reaction mechanism developed by Marc Julia

Table 2. Values calculated through the Density Functional Theory of synthesizing vitamin A acetate by Julia's method 1

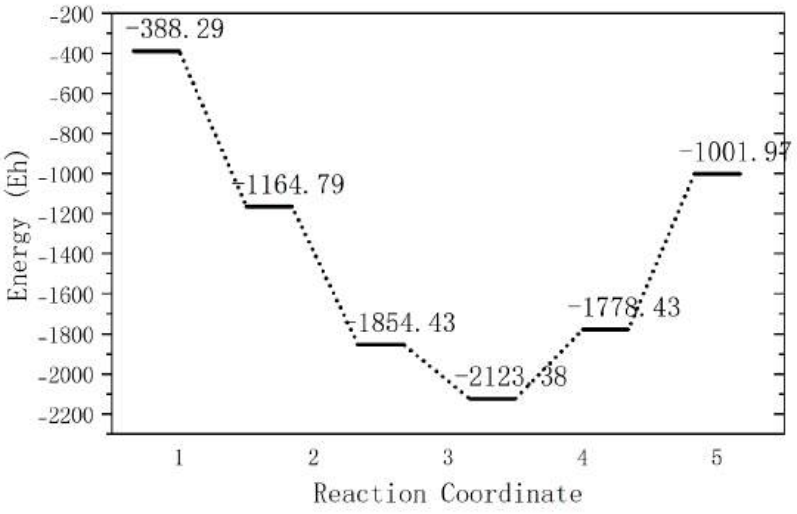
Compound name	***SCF Energy (Hartree)***	***HOMO energy (Hartree)***	***Dipole magnitude (Debye)***
β - Ionone	-579.46179	-0.218577	3.212844591
No. 14	-657.59455	-0.195693	1.13266856
No.16	-1357.7324	-0.20089	4.545812364
No. 19	-1701.3313	-0.203717	5.332725768
No. 7	-849.02156	-0.195027	5.997955331
Vitamin A Acetate	-1001.974713	-0.182186	1.794558566

Table 3 Values calculated through the Density Functional Theory of synthesizing vitamin A acetate by Julia's method 2

Compound name	***SCF Energy (Hartree)***	***HOMO energy (Hartree)***	***Dipole magnitude (Debye)***
β - Ionone	-579.46179	-0.218577	3.212844591
No. 14	-657.59455	-0.195693	1.13266856
No.16	-1357.7324	-0.20089	4.545812364
No. 20	-1778.4217	-0.204485	5.08082873
Vitamin A Acetate	-1001.974713	-0.182186	1.794558566

Kuraray synthesis of vitamin A acetate

Kuraray's synthesis includes a combination of two C_{10} building blocks. Sulfone **29** and aldehyde **30**, were produced from myrcene **27** or linalool **28**. When these two substances are combined, β-hydroxy-sulfone **31** is created, which is then transformed into vitamin A acetate. The mechanism is shown in Scheme 9. The schematic representation of the energy of the mechanism is presented in Scheme 13. The values calculated are presented in Table 4.

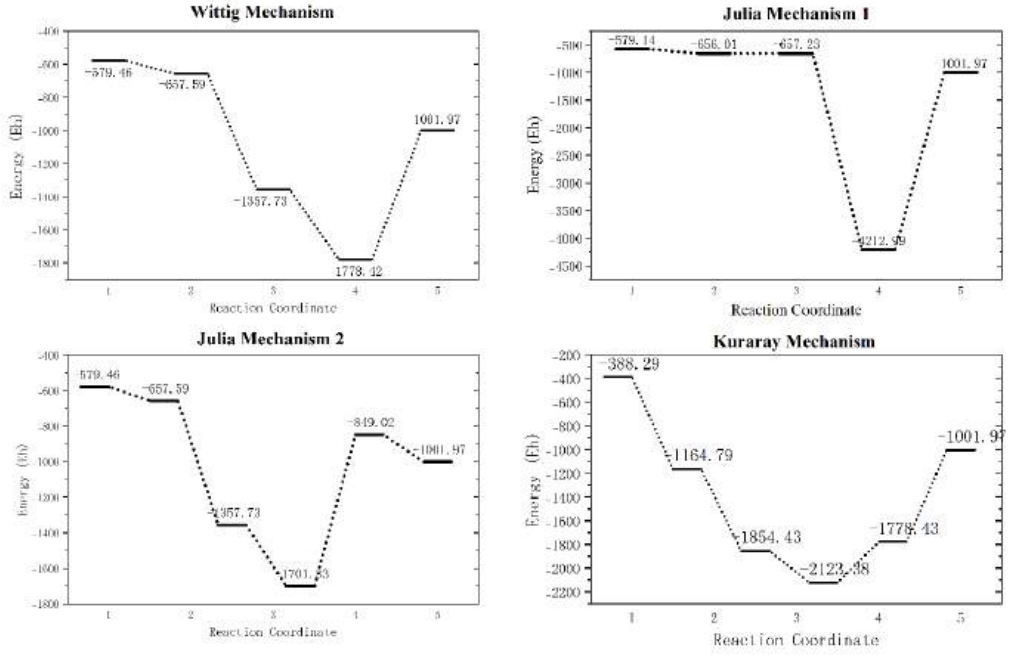


Scheme 13. Schematic representation (energy V.S. reaction coordinate) of synthesis of vitamin A acetate developed by Kuraray

Table 3 values calculated through the Density Functional Theory of synthesizing vitamin A acetate by Kuraray

Compound name	***SCF Energy (Hartree)***	***HOMO energy (Hartree)***	***Dipole magnitude (Debye)***
No. 27	-388.28719	-0.219925	0.39529421
No. 29	-1164.791914	-0.220583	4.669811213
No. 31	-1854.43498	-0.22186	4.70231496
No. 32	-2123.384953	-0.208417	6.409372827
No. 33	-1778.43284	-0.213869	3.843230534
Vitamin A Acetate	-1001.974713	-0.182186	1.794558566

Overall, the synthesis of vitamin A acetate through the Wittig approach is the one that has the highest energy change, whereas the first approach developed by Marc Julia has the lowest energy change. The schematic representation of all reaction mechanisms is shown in scheme 14.



Scheme 14. Schematic representation (energy V.S. reaction coordinate) of all reaction mechanisms of synthesizing vitamin A acetate

Discussion and conclusion

This DFT study compared three synthesis pathways for Vitamin A Acetate, which includes the Wittig approach, Julia Chemistry, and Kuraray's approach. Through the calculations, energy change of all three mechanisms could be compared.

The Wittig approach, which involves the elimination of a C15 building block with a C5 moiety, resulted in the highest energy changes throughout all three reaction mechanisms, which should be one of the reasons why this method is not widely used as other reaction mechanisms. In contrast, the first approach by Marc Julia exhibited the lowest energy changes, making it the most energetically efficient pathway among all three. These results aligned with its continued use in industrial processes, especially in large-scale production of vitamin A. The second route developed by Julia, although less energetically favourable than the first one, still performs comparably to other synthesis methods. Kuraray's $C_{10} + C_{10}$ approach, while innovative, presented an intermediate energy change between the Wittig and Julia methods. Thus, in terms of energy, Julia's method of synthesizing vitamin A is the most efficient.

However, the experiment still has plenty of limitations. Other factors such as scalability, raw material availability, and industrial preferences must be considered when selecting the optimal synthesis route for Vitamin A acetate. The results from this DFT study only give a reference to evaluating the synthetic pathways and for further optimization in industrial applications.

Appendix

Appendix 1: Python code for generating chemical structures from SMILES

```
!pip install --prefer-binary pyscf
!pip install pyberny
#!pip install geometric
!pip install fortecubeview pythreejs
!pip install --upgrade traitlets
!pip install rdkit
!pip install py3Dmol


from google.colab import output
output.enable_custom_widget_manager()
import pathlib
# RDKit imports:
from rdkit import Chem
from rdkit.Chem import (
    AllChem,
    rdCoordGen,
)
from rdkit.Chem.Draw import IPythonConsole
IPythonConsole.ipython_useSVG = True  # Use higher quality images for molecules
# For visualization of molecules and orbitals:
import py3Dmol
import fortecubeview
# pyscf imports:
from pyscf import gto, scf, lo, tools
# For plotting
import matplotlib
from matplotlib import pyplot as plt
import seaborn as sns
%matplotlib inline
sns.set_theme(style="ticks", context="talk", palette="muted")
# For numerics:
import numpy as np
import pandas as pd
pd.options.display.float_format = "{:,.3f}".format


molecule_name = "Bicarbonate"
molecule = Chem.MolFromSmiles("")  # Generate the molecule from smiles
molecule


def get_xyz(molecule, optimize=False):
    """Get xyz-coordinates for the molecule"""
    mol = Chem.Mol(molecule)
    mol = AllChem.AddHs(mol, addCoords=True)
    AllChem.EmbedMolecule(mol)
    if optimize:  # Optimize the molecules with the MM force field:
        AllChem.MMFFOptimizeMolecule(mol)
    xyz = []
    for lines in Chem.MolToXYZBlock(mol).split("\n")[2:]:
        strip = lines.strip()
        if strip:
            xyz.append(strip)
    xyz = "\n".join(xyz)
    return mol, xyz


molecule3d, xyz = get_xyz(molecule)
print(xyz)


view = py3Dmol.view(
    data=Chem.MolToMolBlock(molecule3d),
    style={"stick": {}, "sphere": {"scale": 0.3}},
    width=300,
    height=300,
)
view.zoomTo()
```

Works Cited

Chapman, M. Shane. “Vitamin A: History, Current Uses, and Controversies.” *Seminars in Cutaneous Medicine and Surgery* 31, no. 1 (March 2012): 11–16. https://doi.org/10.1016/j.sder.2011.11.009.

Stevens, Gretchen A, James E Bennett, Quentin Hennocq, Yuan Lu, Luz Maria De-Regil, Lisa Rogers, Goodarz Danaei, et al. “Trends and Mortality Effects of Vitamin a Deficiency in Children in 138 Low-Income and Middle-Income Countries between 1991 and 2013: A Pooled Analysis of Population-Based Surveys.” *The Lancet Global Health* 3, no. 9 (September 2015): e528–36. https://doi.org/10.1016/s2214-109x(15)00039-x.

Semba, Richard D. “On the ‘Discovery’ of Vitamin A.” *Annals of Nutrition & Metabolism* 61, no. 3 (November 26, 2012): 192–98. https://doi.org/10.1159/000343124.

Rosenfeld, Louis. “Vitamine—Vitamin. The Early Years of Discovery.” *Clinical Chemistry* 43, no. 4 (April 1997): 680–85. https://web.archive.org/web/20160604072512/http://www.clinchem.org/content/43/4/680.full.

Drummond, Jack Cecil. “The Nomenclature of the So-Called Accessory Food Factors (Vitamins).” *Biochemical Journal* 14, no. 5 (October 1, 1920): 660–60. https://doi.org/10.1042/bj0140660.

Karrer, P, R Morf, and K. Schöpp. “Zur Kenntnis Des Vitamins‐a Aus Fischtranen II.” *Helvetica Chimica Acta* 14, no. 6 (December 1, 1931): 1431–36. https://doi.org/10.1002/hlca.19310140622.

Holmes, Harry N, and Ruth E Corbet. “The Isolation of Crystalline Vitamin a[1].” *Journal of the American Chemical Society* 59, no. 10 (October 1, 1937): 2042–47. https://doi.org/10.1021/ja01289a075.

ARENS, J. F., and D. A. VAN DORP. “Synthesis of Some Compounds Possessing Vitamin a Activity.” *Nature* 157, no. 3981 (February 1946): 190–91. https://doi.org/10.1038/157190a0.

VAN DORP, D. A., and J. F. ARENS. “Synthesis of Vitamin a Aldehyde.” *Nature* 160, no. 4058 (August 1947): 189–89. https://doi.org/10.1038/160189a0.

Parker, Gemma L., Laura K. Smith, and Ian R. Baxendale. “Development of the Industrial Synthesis of Vitamin A.” *Tetrahedron* 72, no. 13 (March 2016): 1645–52. https://doi.org/10.1016/j.tet.2016.02.029.

Arens, J. F., and D. A. van Dorp. “Synthesis of Vitamin A.” *Recueil Des Travaux Chimiques Des Pays-Bas* 68, no. 7 (September 2, 2010): 604–8. https://doi.org/10.1002/recl.19490680703.

Isler, Otto, Wolfgang Huber, A Ronco, and Markus Kofler. “Synthese Des Vitamin A” 30, no. 6 (October 15, 1947): 1911–27. https://doi.org/10.1002/hlca.19470300666.

Wittig, Georg, and Georg Geissler. “Zur Reaktionsweise des Pentaphenyl-Phosphors Und Einiger Derivate.” *Justus Liebigs Annalen, Chemie* 580, no. 1 (March 27, 1953): 44–57. https://doi.org/10.1002/jlac.19535800107.

Sarnecki, Wilhelm, and Horst Pommer. "Verfahren zur Herstellung von (ª‰-Ion Yliden Ethyl)-triaryl phosphonium halogeniden."

ARNOULD, D., P. CHABARDES, G. FARGE, and M. JULIA. “ChemInform Abstract: Organic Syntheses with Sulfones. Part 23. A Laboratory Procedure for the Synthesis of Retinoic Acid.” *Chemischer Informationsdienst* 16, no. 35 (September 3, 1985). https://doi.org/10.1002/chin.198535340.

Gould, R. G., and A. F. Thompson. “The Synthesis of Certain Unsaturated Compounds from Beta-Ionone and Tetrahydro Ionone 1[,2,3].” *Journal of the American Chemical Society* 57, no. 2 (February 1935): 340–45. https://doi.org/10.1021/ja01305a032.

Pommer, H., and A. Nürrenbach. “Industrial Synthesis of Terpene Compounds.” *Pure and Applied Chemistry* 43, no. 3-4 (January 1, 1975): 527–51. https://doi.org/10.1351/pac197543030527.

Waddell, Walter H., and Kohji Chihara. “Activation Barriers for the Trans .Fwdarw. Cis Photoisomerization of All-Trans-Retinal.” *Journal of the American Chemical Society* 103, no. 24 (December 1981): 7389–90. https://doi.org/10.1021/ja00414a083.

Fischer, M., D. Horn, F. Feichtmayr, W. Wiersdorff, and A. Nuerrenbach. "Isomerization of vitamin A compounds and their derivatives." U.S. Patent 3,838,029, issued September 24, 1974.

Schleich, K., and H. Stoller. *Verfahren zur Herstellung von Vitamin-A-Acetat*. Patent DE2733231, 1978.

Schaefer, Bernd, Wolfgang Siegel, Steffen Tschirschwitz, Till Brüggemann, Marius Sorin Pulbere, and Florian Buchbender. "Isomerization of polyunsaturated non-aromatic compounds." U.S. Patent Application 17/616,911, filed October 6, 2022.

Mori, Toshiki, and Kensuke Nagashima. *Production of 11-Trans-Vitamin A Compounds*. Japan Patent JPH11147868A, filed November 13, 1997, and published June 2, 1999.

Bonrath, Werner, Bo Gao, Peter Houston, Tom McClymont, Marc-André Müller, Christian Schäfer, Christiane Schweiggert, Jan Schütz, and Jonathan Medlock. “75 Years of Vitamin a Production: A Historical and Scientific Overview of the Development of New Methodologies in Chemistry, Formulation, and Biotechnology.” *Organic Process Research & Development* 27, no. 9 (August 14, 2023). https://doi.org/10.1021/acs.oprd.3c00161.

Julia, M. "Sulfone intermediates for the synthesis of vitamin a." U.S. Patent 3,781,313, issued December 25, 1973.

Julia, Mare. *Neue Zwischenprodukte der Vitamin-A-Synthese*. Germany Patent DE2202689A1, filed January 20, 1972, and published August 3, 1972.

Decor, Jean-Pierre. "Alpha-halogen acetale von aethylenischen aldehyden und verfahren zu ihrer herstellung." German Democratic Republic Patent DD128609A5, filed February 24, 1977, and published November 30, 1977.

Chabardes, Pierre, Marc Julia, and Albert Menet. *Neue von 1,5-Dimethylhexadien(1,5)-ylen Abgeleitete Sulfone*. Germany Patent DE2305267A1. Filed February 2, 1972. Accessed November 24, 2024.

Chabardes, Pierre, Marc Julia, and Albert Menet. "Neue sekundaere und tertiaire sulfone." 19732355898, filed May 16, 1974.

Grosselin, Jean Michel, C. Mercier, G. Allmang, and F. Grass. "Selective Hydrogenation of α,β-Unsaturated Aldehydes in Aqueous Organic Two-Phase Solvent Systems Using Ruthenium or Rhodium Complexes of Sulfonated Phosphines." *Organometallics* 10, no. 7 (July 1, 1991): 2126–2133. https://doi.org/10.1021/om00053a014.

Fischli, Albert, and Hans Mayer. "Carotenoid Synthesis über Sulfone; Synthese von Apocarotinoiden und Torularhodinester" Helvetica Chimica Acta 58, no. 6 (July 16, 1975): 1584–90. https://doi.org/10.1002/hlca.19750580611.

Masanao, Matsui. "Method for preparing polyene carboxylic acid derivatives." U.S. Patent 2,951,853, issued September 6, 1960.

Matsui, Masanao, Shigeru Okano, Kyohei Yamashita, Masateru Miyano, Seiichi Kitamura, Akio Kobayashi, Tomiichi Sato, and Ryuzo Mikami. "SYNTHETIC STUDIES on VITAMIN A." *The Journal of Vitaminology* 4, no. 3 (January 1, 1958): 178–89. https://doi.org/10.5925/jnsv1954.4.178.

Huisman, H. O., & Smit, A. (1958).

Huisman, H. O., A. Smit, P. H. van Leeuwen, and J. H. van Rij. "Investigations in the Vitamin a Series. III. Rearrangement of the Retro-System to the Normal System of Conjugated Double Bonds in the Vitamin a Series." *Recueil Des Travaux Chimiques Des Pays-Bas* 75, no. 9 (September 2, 2010): 977–1006. https://doi.org/10.1002/recl.19560750902.

Humphlett, Wilbert J. "Method of making vitamin a and intermediates formed thereby." U.S. Patent 2,676,992, issued April 27, 1954.

Humphlett, Wilbert J., and Donald M. Burness. "Method of making vitamin A and intermediates formed thereby." U.S. Patent 2,676,990, issued April 27, 1954.

Lagerweij, Gerrit J., Cornelis Bakker, and Monique EA De Bruin-Van Der. "Use of an allylchloride for preparing an aldehyde." U.S. Patent 5,196,608, issued March 23, 1993.

Otera, Junzo, Tadakatsu Mandai, and Mikio Kawada. "Process for producing vitamin A or its carboxylic acid esters, and intermediate compounds useful for the process." U.S. Patent 4,825,006, issued April 25, 1989.

Junzo Otera, Hiromitsu Misawa, Takashi Onishi, Shigeaki Suzuki, and Yoshiji Fujita. "Stereocontrolled Synthesis of Vitamin a through a Double Elimination Reaction. A Novel Convergent C10 + C10 Route." *The Journal of Organic Chemistry* 51, no. 20 (October 1, 1986): 3834–38. https://doi.org/10.1021/jo00370a017.

Chemla, F., M. Julia, and D. Uguen. "A stereoselective C10+ C10 route to retinal." Bulletin de la Société chimique de France 130, no. 2 (1993): 200-205.

Ladd, Elbert C. "Preparation of aromatic sulfones." U.S. Patent 2,573,580, issued October 30, 1951.

Mori, Toshiki, Shigeaki Suzuki, Takashi Onishi, and Kazuo Yamamoto. "Process for producing α, β-unsaturated aldehyde." U.S. Patent 5,053,552, issued October 1, 1991.

Suzuki, Shigeaki, Toshiki Mori, Takashi Onishi, and Yoshiji Fujita. "Process for preparing chlorinated olefins." U.S. Patent 4,827,056, issued May 2, 1989.

Otera, Junzo, Shigeaki Suzuki, Takashi Onishi, and Yoshiji Fujita. "Process for preparing α, β-unsaturated aldehydes." U.S. Patent 4,745,229, issued May 17, 1988.

Jones, Robert O. "Density functional theory: Its origins, rise to prominence, and future." *Reviews of modern physics* 87, no. 3 (2015): 897-923.

Townsend, Piers A., and Matthew N. Grayson. "Density functional theory in the prediction of mutagenicity: a perspective." *Chemical research in toxicology* 34, no. 2 (2020): 179-188.

Echenique, Pablo, and José Luis Alonso. "A mathematical and computational review of Hartree–Fock SCF methods in quantum chemistry." *Molecular Physics* 105, no. 23-24 (2007): 3057-3098.

Lee, Chengteh, Weitao Yang, and Robert G. Parr. "Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density." *Physical Review B* 37, no. 2 (January 15, 1988): 785–89. https://doi.org/10.1103/physrevb.37.785.

Neese, Frank. "Software Update: The ORCA Program System—Version 5.0." *WIREs Computational Molecular Science*, March 7, 2022. https://doi.org/10.1002/wcms.1606.

Neese, Frank. "An Improvement of the Resolution of the Identity Approximation for the Formation of the Coulomb Matrix." *Journal of Computational Chemistry* 24, no. 14 (September 3, 2003): 1740–47. https://doi.org/10.1002/jcc.10318.

Neese, Frank, Frank Wennmohs, Andreas Hansen, and Ute Becker. "Efficient, Approximate and Parallel Hartree–Fock and Hybrid DFT Calculations. A 'Chain-of-Spheres' Algorithm for the Hartree–Fock Exchange." *Chemical Physics* 356, no. 1-3 (February 2009): 98–109. https://doi.org/10.1016/j.chemphys.2008.10.036.

Bykov, Dmytro, Taras Petrenko, Róbert Izsák, Simone Kossmann, Ute Becker, Edward Valeev, and Frank Neese. "Efficient Implementation of the Analytic Second Derivatives of Hartree–Fock and Hybrid DFT Energies: A Detailed Analysis of Different Approximations." *Molecular Physics* 113, no. 13-14 (March 25, 2015): 1961–77. https://doi.org/10.1080/00268976.2015.1025114.

Stoychev, Georgi L., Alexander A. Auer, and Frank Neese. "Automatic Generation of Auxiliary Basis Sets." *Journal of Chemical Theory and Computation* 13, no. 2 (January 10, 2017): 554–62. https://doi.org/10.1021/acs.jctc.6b01041.

Helmich-Paris, Benjamin, Bernardo de Souza, Frank Neese, and Róbert Izsák. "An improved chain of spheres for exchange algorithm." *The Journal of Chemical Physics* 155, no. 10 (2021).

Neese, Frank. "The SHARK integral generation and digestion system." Journal of Computational Chemistry 44, no. 3 (2023): 381-396.

Caldeweyher, Eike, Christoph Bannwarth, and Stefan Grimme. "Extension of the D3 dispersion coefficient model." *The Journal of chemical physics* 147, no. 3 (2017).

Caldeweyher, Eike, Sebastian Ehlert, Andreas Hansen, Hagen Neugebauer, Sebastian Spicher, Christoph Bannwarth, and Stefan Grimme. "A generally applicable atomic-charge dependent London dispersion correction." *The Journal of chemical physics* 150, no. 15 (2019).

Caldeweyher, Eike, Jan-Michael Mewes, Sebastian Ehlert, and Stefan Grimme. "Extension and evaluation of the D4 London-dispersion model for periodic systems." *Physical Chemistry Chemical Physics* 22, no. 16 (2020): 8499-8512.

Nanotech-enabled Cancer Therapies: Advances and Future Prospects
By Sowmithra Pradheepan

Introduction

Cancer is a complex and multifaceted disease, and one of the leading causes of mortality worldwide. According to WHO estimates, in 2022 at least 20 million people were newly diagnosed with some type of cancer, and 9.7 million died due to complications from cancer. Moreover, the organization projects that roughly 20% of the world's population will develop cancer at some point in their lives (World Health Organization). There is an urgent need for innovative approaches to diagnosis and treatment of various types of cancer (Rina et al.; Sheikh et al.)

Nanotechnology is emerging as a transformative force in cancer diagnosis and therapy. By enhancing drug delivery mechanisms and increasing treatment specificity, it offers a promising avenue for overcoming limitations in traditional cancer treatments (Stephen et al., 2020). These advancements herald a crucial shift toward personalized therapies and more effective management of cancer's global burden.

Introducing Nanotechnology

Nanotechnology, or nanotech for short, is an interdisciplinary domain integrating physics, chemistry, biology, materials science, and engineering. Nanotech operates at the nanoscale—1 to 100 nanometers. At that scale, materials exhibit very different properties than at bulk scales, such as millimeter or larger, and can be used for revolutionary applications in fields such as medicine, environmental sciences, and material engineering. The U.S. National Nanotechnology Initiative (NNI) plays a pivotal role in advancing this technology, allocating $1.3 billion annually to its research and development (Bozeman et al., 2007). Similarly, Israel's National Nanotechnology Initiative (INNI) has provided substantial funding, underscoring global interest in nanotech innovation.

Specifically In the field of medicine, nanotechnology enables applications like artificial blood cells, DNA repair, and cellular detoxification, with far-reaching implications for treating cancer and other diseases. This paper delves into the impact of recent advancements in nanotechnology on cancer diagnosis and therapy, highlighting both achievements and challenges.

Current Treatment Methods

Before delving into nanotech-based cancer treatment methods, it is important to understand current therapies against cancer, where their shortcomings lie, and how nanotechnology can be harnessed to bridge these gaps.

Chemotherapy

The most common non-invasive treatment used against cancers is chemotherapy. Traditionally, chemotherapy refers to the use of cytotoxic drugs that are particularly harmful to

rapidly growing cells, as malignant cancers tend to be. They interfere with the process of cell division and induce cell death in the cancerous cells (Johnstone et al.). However, since traditional chemotherapy is not very specific, these drugs will also interfere in the replication and maintenance of healthy tissues and cause unpleasant side effects

While chemotherapy remains a mainstay, issues such as cancers developing multidrug resistance (MDR) and severe side effects limit its efficacy. Common side effects, such as nausea, fatigue, and immunosuppression, remain a significant barrier to patient quality of life.

Surgery

Surgery remains one of the most critical treatment modalities for cancer, particularly in the management of cancers with well-defined tumors. Its utility lies in its ability to physically remove tumor masses, often offering the highest chance of cure in localized cancers. Advances in surgical techniques and technologies have further optimized its effectiveness and minimized associated risks. (Bennardo, Bennardo et al.) (Oneda et al.)

Modern surgical oncology has benefited greatly from technological advancements, especially in imaging, image processing, and robotics. Techniques such as laparoscopic and robotic-assisted surgeries reduce complications, shorten hospital stays, and preserve healthy tissues. Innovations like intraoperative imaging allow surgeons to distinguish cancerous tissue from healthy tissue, improving precision. Meanwhile, AI-driven imaging and real-time analytics aid in surgical planning, optimizing outcomes by predicting the likelihood of recurrence and personalizing approaches. Emerging technologies such as the "intelligent knife" use mass spectrometry to identify cancerous tissues during surgery, potentially reducing recurrence rates. (Montagne et al.; Keelan et al.)

Challenges in Surgery

Surgery, though often curative, is not devoid of challenges. Risks include incomplete resection, damage to adjacent organs, and post-surgical complications. For cancers like pancreatic cancer, surgical success is heavily influenced by early diagnosis and the feasibility of complete tumor removal. Furthermore, the emergence of systemic therapies like immunotherapy has led to a paradigm shift, where surgery is increasingly integrated into multimodal treatment strategies.

Immunotherapy

Immunotherapy is a groundbreaking approach in cancer treatment, leveraging the body’s immune system to detect and destroy cancer cells. Unlike traditional therapies such as chemotherapy and radiotherapy, which directly target cancer cells but may harm normal tissues, immunotherapy focuses on enhancing the immune system's ability to fight tumors. This therapeutic modality includes checkpoint inhibitors, monoclonal antibodies, cytokine therapies, and cancer vaccines, each playing a vital role in improving patient outcomes.

Immunotherapy works by disrupting the mechanisms cancer cells use to evade immune detection. For example, immune checkpoint inhibitors such as pembrolizumab and nivolumab target checkpoint proteins like PD-1 and CTLA-4, which cancer cells use to suppress the ability of immune cells to detect and kill them. By blocking these checkpoints, T-cells regain the ability to attack cancer cells. (Ling et al.)

These drugs have revolutionized cancer treatment by enabling durable responses, especially in cancers with high mutational loads such as melanoma and NSCLC. For instance, nivolumab has demonstrated improved recurrence-free survival in melanoma patients compared to earlier treatments like ipilimumab

Recent clinical trials have explored neoadjuvant (pre-surgery) checkpoint blockade to prime systemic immunity. Early findings indicate higher tumor-specific T-cell responses and better outcomes in cancers such as melanoma and colorectal carcinoma (Ling et al.)

Nano-integrated Treatment Methods

Cancer Progress Report, Mayo Clinic News Network Nanotechnology offers solutions through nanomedicine, improving drug delivery and targeting resistant cells, though clinical translation remains a challenge (Bukowski et al.) Zhu et al.).

Nanotech and Cancer

Nanotechnology has revolutionized cancer diagnostics through the development of nanoparticles conjugated with targeting molecules, enabling precise tumor identification and interaction. Innovations such as metal-based nanomaterials, quantum dots, magnetic nanoparticles, and lab-on-chip devices leverage the tumor microenvironment to enhance specificity (Stephen et al.).

Therapeutic advancements include precision medicines, immunotherapies, epigenetic interventions, and RNA-based therapies targeting specific molecular pathways (Shifana et al.). Nanogel-based immunotherapy exemplifies this, offering efficient delivery systems for drugs and genetic material, but challenges such as immune-related adverse events persist (Ma et al.) Moreover, pulmonary nanomedicine delivery and gas therapies have emerged as groundbreaking approaches in addressing diseases like lung cancer and exploring combined treatment modalities (A, Han (Ji et al.).

Nanotechnology enhances cancer photo therapies by using light-responsive nanoparticles. Photodynamic therapy (PDT) uses photosensitizers that, upon light activation, generate reactive oxygen species to destroy cancer cells. Photothermal therapy (PTT), on the other hand, employs nanoparticles that convert light energy into heat to target and kill malignant cells. These therapies are gaining traction due to their ability to selectively target tumors while sparing healthy tissue (Shabnum et al.)

It is also revolutionizing cancer diagnostics. Quantum dots and gold nanoparticles enhance imaging techniques like MRI and CT scans by providing high-resolution insights into tumor morphology. Theranostics—a combination of diagnostics and therapy—leverages

nanoparticles to diagnose cancer and deliver treatments simultaneously, streamlining patient care (Dutta Gupta et al.)

Diagnostics

Nanotechnology is revolutionizing cancer diagnostics by providing highly sensitive and precise tools for tumor detection. Quantum dots and gold nanoparticles are at the forefront of this innovation. (Shabnum et al.)

Quantum Dots: Quantum dots are nanometer-sized semiconductor particles that exhibit unique optical properties, such as bright fluorescence and tunable emission wavelengths, which enhance imaging techniques like MRI and CT scans. Quantum dots offer high-resolution insights into tumor morphology, enabling early detection and improved diagnostic accuracy.

Gold Nanoparticles**:** Gold nanoparticles are versatile agents in diagnostics due to their ability to scatter and absorb light effectively. They are used in imaging techniques and biosensors, enhancing the detection of cancer biomarkers and providing detailed visualizations of cancerous tissues.

Targeted Drug Delivery

Nanotechnology is advancing chemotherapy by enabling more precise drug delivery mechanisms, minimizing side effects, and improving therapeutic outcomes. (Williams et al.)

Liposomal Drug Delivery**:** Liposomes are spherical vesicles composed of lipid bilayers that encapsulate chemotherapeutic agents. These structures improve the stability and bioavailability of drugs while reducing toxicity by targeting cancer cells specifically.

Nanoparticle Envelopes**:** Nanoparticles, such as polymeric and metallic nanoparticles, serve as efficient drug carriers. Their small size allows them to penetrate tumors through the enhanced permeability and retention (EPR) effect. Functionalized nanoparticles can be engineered to release drugs in response to specific stimuli like pH or temperature changes within the tumor microenvironment.

Phytotherapies

Nanotechnology enhances cancer phototherapies by leveraging light-responsive nanoparticles to selectively target tumors (Shabnum et al.)

Photodynamic Therapy (PDT)**:** PDT utilizes photosensitizers that, when activated by light, generate reactive oxygen species (ROS) to destroy cancer cells. Nanoparticles improve the delivery and activation of photosensitizers, increasing the efficacy of this therapy while minimizing damage to surrounding healthy tissues.

Photothermal Therapy (PTT)**:** PTT involves nanoparticles that convert light energy into heat to target and kill malignant cells. Gold nanoparticles, carbon nanotubes, and other nanomaterials are commonly used to achieve localized heating, effectively destroying tumors with minimal side effects.

Immunotherapies

Nanotechnology is transforming immunotherapy by enhancing the activation and delivery of immune-modulating agents. (Li et al.)

Nanoparticles can be engineered to deliver immune checkpoint inhibitors, cytokines, or antigens directly to immune cells or tumor sites. This targeted approach reduces systemic toxicity and improves the efficacy of immunotherapeutic interventions.

Nanovaccines and nanoparticle-based adjuvants are being developed to stimulate robust anti-tumor immune responses, offering new avenues for cancer immunotherapy.

Vaccines

Vaccines in cancer treatment, often referred to as cancer vaccines, represent a growing and innovative area within oncology. They aim to stimulate the immune system to recognize and combat cancer cells by targeting tumor-specific antigens (TSAs) or tumor-associated antigens (TAAs). These antigens can either be unique to individual tumors or shared across specific cancer types, making vaccines a versatile therapeutic option.

Cancer vaccines work by presenting TSAs or TAAs to the immune system, enabling recognition and destruction of cancer cells. For instance, mRNA vaccines encode antigens that, when introduced into the body, produce proteins recognized as foreign, stimulating T-cell activation and memory development. Advances in delivery systems, such as lipid nanoparticles, have further improved their stability and efficacy. (Chehelgerdi and Chehelgerdi)

Nanotechnology enhances vaccine delivery, improving antigen presentation and immune activation, which shows potential in cancer immunotherapy. Sipuleucel-T is one such vaccine designed to elicit an immune response against TSAs associated with prostate cancer. It has demonstrated the potential to enhance survival while being well-tolerated by patients (Ling et al.)

In addition, nanoparticle-based adjuvants improve the efficacy of vaccines by boosting the immune response and ensuring prolonged antigen release, increasing the likelihood of tumor recognition and destruction

Despite these advancements, challenges remain. Nanoparticles must navigate complex biological environments, which can affect their efficacy. Issues like biodegradability, potential toxicity, and scalability of production are critical hurdles. However, research is continually addressing these concerns, refining the safety and efficiency of nanotechnology-based solutions.

Latest Research

Recent advancements in nanotechnology include the use of gas-based nanomedicine for targeted tumor treatment and innovations in mRNA vaccine technology, supported by lipid

nanoparticle-based systems (Ji et al.; Huang et al.). Novel nanoscale covalent–organic frameworks (COFs) have shown promise in integrating imaging and therapeutic agents, though challenges in biocompatibility remain (Kaur et al.). Colon-targeted drug release methods, such as enzyme-sensitive polymers, are also gaining traction in enhancing selectivity and cellular uptake (Bhirud et al.)

The integration of artificial intelligence (AI) with nanotechnology offers exciting prospects. AI-driven algorithms can optimize nanoparticle design for individualized cancer treatments. Furthermore, advances in biodegradable and bio-responsive nanoparticles aim to mitigate long-term toxicity risks.

Comparison Between Traditional and Nano-integrated Treatment Methods

The evolution of cancer treatment from traditional methods like surgery, chemotherapy, and radiation therapy to nano-integrated therapies reflects significant advancements in precision, efficiency, and patient outcomes. This shift has introduced opportunities for overcoming the limitations of conventional approaches, yet both modalities come with distinct benefits and challenges.

While traditional treatments often face limitations like non-specificity and systemic toxicity, nano-enabled methods improve targeting and efficacy. However, nanotechnology's high development costs and limited clinical trials present barriers to widespread adoption.

Aspect	**Traditional Methods**	**Nano-integrated Methods**
Specificity	Limited; affects healthy tissues.	High; targets tumor-specific biomarkers.
Systemic effects	Significant systemic toxicity.	Reduced; localized delivery lowers off-target effects.
Efficacy in Metastases	Limited; challenges in targeting systemic spread.	Effective; nanoparticles can target multiple sites.
Cost	Relatively lower upfront costs.	Higher due to research, production, and development.
Development Stage	Well-established; broadly available.	Emerging; requires more clinical trials for validation.
Regulatory Challenges	Minimal; long history of approval pathways.	Significant; regulatory frameworks for nanomedicine are still evolving .

While traditional treatments remain the foundation of cancer treatments, nano-integrated methods offer revolutionary potential by addressing the shortcomings of conventional approaches. The integration of nanotechnology enhances targeting precision, reduces toxicity, and enables innovative strategies like theranostics and immunotherapy. However, the scalability,

cost, and regulatory barriers associated with nanomedicine must be addressed to achieve widespread adoption. (Dutta Gupta et al.)

Challenges, Open Questions, and Opportunities

Despite remarkable progress, nanotech-enabled cancer therapies face challenges such as limited efficacy, regulatory hurdles, and cytotoxicity concerns. However, they present opportunities to refine targeted delivery systems, enhance combination therapies, and develop precision medicines tailored to individual profiles (Ma et al.; Ji et al.)

Limitations in Efficacy

While nanotechnology offers precision and targeted delivery, its efficacy is often limited by challenges such as uneven nanoparticle distribution within tumors. For example, the enhanced permeability and retention (EPR) effect, which facilitates nanoparticle accumulation in tumors, is inconsistent across different tumor types and patients. This variability can reduce the therapeutic impact, necessitating personalized approaches to improve outcomes.

Regulatory Hurdles

The approval of nanotechnology-based therapies faces stringent regulatory challenges due to their complexity and novelty. Regulatory standards aim to ensure the safety, efficacy, and quality of nanomedicines. The unique properties of nanoparticles, such as their size, surface chemistry, and biological interactions, require specialized evaluation methods that differ from traditional pharmaceuticals.

Nanotech therapies must meet guidelines set by agencies like the FDA in the United States and EMA in Europe, including rigorous preclinical testing for toxicity, pharmacokinetics, and biodistribution. Additionally, manufacturing processes must comply with Good Manufacturing Practices (GMP) to ensure consistency and safety.

Adverse Effects

Nanotechnology-enabled therapies can cause unintended adverse effects, which limit their widespread adoption.

Cytotoxicity**:** Some nanoparticles, such as quantum dots or metallic nanoparticles, release toxic ions or generate reactive oxygen species that can damage healthy cells. For instance, silver nanoparticles have been shown to induce oxidative stress, leading to cell death.

Allergic Reactions**:** Certain nanoparticle formulations may trigger immune responses, causing inflammation or hypersensitivity. For example, liposomal formulations can sometimes activate the complement system, leading to allergic-like reactions.

Addressing these challenges through innovative research and development is crucial to maximizing the potential of nanotechnology in cancer therapy

Conclusion

The integration of nanotechnology into cancer treatment heralds a transformative era in oncology. While challenges remain, the advancements discussed in this paper underscore its potential to revolutionize diagnosis and therapy, paving the way for a future of precise, personalized, and minimally invasive cancer care.

Works Cited

A, Han et al., 2023. Pulmonary delivery of nanoparticles: Challenges and opportunities in lung cancer therapy. Journal of Nanomedicine Research.

Bennardo, Bennardo et al., 2021**.** Local chemotherapy as an alternative for squamous cell carcinoma: Clinical outcomes and future perspectives. Journal of Oncology Advances.

Bhirud, Darshan, et al. "Bioengineered Carbohydrate Polymers for Colon-Specific Drug Release: Current Trends and Future Prospects." Journal of Biomedical Materials Research Part A, 2024.

Bozeman et al., 2007. Nanotechnology research and its implications in cancer and other industries. National Nanotechnology Initiative Reports.

Bukowski, Karol, et al. "Mechanisms of Multidrug Resistance in Cancer Chemotherapy." International Journal of Molecular Sciences, vol. 21, no. 9, 2020, p. 3233.

Chehelgerdi, Mohammad, and Matin Chehelgerdi. "The Use of RNA-Based Treatments in the Field of Cancer Immunotherapy." Molecular Cancer, vol. 22, no. 1, 2023, p. 106.

Dutta Gupta, Yashaswi, et al. "Mesoporous Silica Nanotechnology: Promising Advances in Augmenting Cancer Theranostics." Cancer Nanotechnology, vol. 15, no. 1, 2024, p. 9.

Huang, Tao, et al. "Lipid Nanoparticle-Based mRNA Vaccines in Cancers: Current Advances and Future Prospects." Frontiers in Immunology, vol. 13, 2022, p. 922301.

Ji, Peng, et al. "Mechanisms and Application of Gas-Based Anticancer Therapies." Pharmaceuticals, vol. 16, no. 10, 2023, p. 1394.

Johnstone, Ricky W., et al. "Apoptosis: A Link between Cancer Genetics and Chemotherapy." Cell, vol. 108, no. 2, 2002, pp. 153–64.

Kaur, Harjot, et al. "Covalent–Organic Framework-Based Materials in Theranostic Applications: Insights into Their Advantages and Challenges." ACS Omega, vol. 9, no. 6, 2024, pp. 6235–52.

Keelan, Stephen, et al. "Evolving Trends in Surgical Management of Breast Cancer: An Analysis of 30 Years of Practice Changing Papers." Frontiers in Oncology, vol. 11, 2021, p. 622621.

Li, Junwei, et al. "A Triple-Combination Nanotechnology Platform Based on Multifunctional RNA Hydrogel for Lung Cancer Therapy." Science China Chemistry, vol. 63, 2020, pp. 546–53.

Ling, Sia Pei, et al. "Role of Immunotherapy in the Treatment of Cancer: A Systematic Review." Cancers, vol. 14, no. 21, 2022, p. 5205.

Ma, Xianbin, et al. "Bioengineered Nanogels for Cancer Immunotherapy." Chemical Society Reviews, vol. 51, no. 12, 2022, pp. 5136–74.

Montagne, Francois, et al. "The Role of Surgery in Lung Cancer Treatment: Present Indications and Future Perspectives—State of the Art." Cancers, vol. 13, no. 15, 2021, p. 3711.

Oneda, Ester, et al. "Biliary Tract Cancer: Current Medical Treatment Strategies." Cancers, vol. 12, no. 5, 2020, p. 1237.

Rina, Angela, et al. “The Genetic Analysis and Clinical Therapy in Lung Cancer: Current Advances and Future Directions.” Cancers, vol. 16, no. 16, 2024, p. 2882.

Shabnum, S. Sameera, et al. “Advancements in Nanotechnology-Driven Photodynamic and Photothermal Therapies: Mechanistic Insights and Synergistic Approaches for Cancer Treatment.” RSC Advances, vol. 14, no. 52, 2024, pp. 38952–95.

Sheikh, Mahdi, et al. “Current Status and Future Prospects for Esophageal Cancer.” Cancers, vol. 15, no. 3, 2023, p. 765.

Shifana, AS, et al. “A Comprehensive Review on Novel Pathways in Cancer Treatment: Clinical Applications and Future Prospects.” Current Cancer Drug Targets.

Stephen, Suchanti et al., 2020. Nanoparticles in cancer diagnostics and therapy: Tumor microenvironment as a target. Nano Research Journal.

Williams, Patrick A., et al. “AACR Cancer Progress Report 2024: Inspiring Science—Fueling Progress—Revolutionizing Care.” Clinical Cancer Research, vol. 30, no. 19, 2024, pp. 4296–98.

World Health Organization. “Global Cancer Burden Growing, amidst Mounting Need for Services.” World Health Organization, 1 Feb. 2024, https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services.

Zhu, Zhang et al., 2022. Targeted drug delivery using nanocarriers in chemotherapy: Clinical hurdles and future directions. Cancer Nanotechnology Journal.

Beyond Black Holes: The Theoretical Reality of White Holes By Navya

Abstract

White holes, the theoretical time-reversed counterparts of black holes, emerge as solutions to Einstein's field equations in general relativity. While black holes absorb matter and energy, white holes are predicted to expel them, making them impervious to external interactions. This paper delves into the mathematical foundations, physical interpretations, and observational challenges associated with white holes. We explore their potential connections to wormholes, their implications for the second law of thermodynamics, and their role in quantum gravity. Additionally, we discuss their cosmological significance, including the hypothesis that the Big Bang itself may have been a white hole event. Despite being a mathematically consistent solution, no observational evidence for white holes currently exists, posing fundamental questions about their physical plausibility.

Author summary

The concept of white holes, while rooted in general relativity, presents numerous challenges in terms of both physical plausibility and observational evidence. This paper explores their theoretical background, including their connection to black holes and wormholes, and their implications for cosmology. We examine the challenges they pose to thermodynamics and discuss their role in quantum gravity and potential links to the resolution of the black hole information paradox. While no empirical evidence currently supports their existence, white holes remain a valuable area of exploration for future theoretical and observational research.

1. Introduction

The study of black holes has fundamentally reshaped our understanding of gravity and spacetime. However, general relativity also predicts another intriguing possibility: white holes. Derived from the Schwarzschild metric, white holes can be considered the time-reversed analogs of black holes. Unlike black holes, which have an event horizon that prevents anything from escaping, white holes possess an event horizon that prevents anything from entering, allowing only outward motion of matter and radiation.

Despite their theoretical validity, white holes face significant challenges in physical realization. This paper explores their origins in general relativity, their implications for astrophysics and cosmology, and the obstacles they present concerning entropy and stability. Furthermore, we analyze the possibility of detecting white holes and discuss their potential role in quantum gravity and the resolution of the black hole information paradox.

2. Theoretical Background

2.1 Einstein's Field Equations and White Hole Solutions

White holes emerge as solutions to Einstein's field equations, specifically in the Schwarzschild and Kerr metrics. The Schwarzschild metric for a non-rotating black hole is given by:

$$ds^2 = -\left(1 - \frac{2GM}{r}\right) dt^2 + \left(1 - \frac{2GM}{r}\right)^{-1} dr^2 + r^2 d\Omega^2$$

where:

- G is the gravitational constant,
- M is the mass of the object,
- r is the radial coordinate,
- represents the angular components.

For a white hole, this metric remains unchanged, but the solution is interpreted with a reversed time orientation, meaning all geodesics lead outward rather than inward.

2.2 Connection to Wormholes

A compelling implication of white holes is their potential relationship with wormholes. The maximally extended Schwarzschild solution consists of two asymptotically flat regions connected by a non-traversable Einstein-Rosen bridge (wormhole). This connection implies that a black hole in one region may be linked to a white hole in another. The metric describing a non-traversable wormhole is:

$$ds^2 = -c^2 d\tau^2 + d\ell^2 + r^2 d\Omega^2$$

where l represents the proper distance through the wormhole throat.

However, classical general relativity suggests such wormholes are highly unstable and collapse almost instantly. Quantum effects or exotic matter with negative energy density would be required to stabilize them, raising further questions about their physical viability.

2.3 Thermodynamic Constraints

A major challenge to the existence of white holes is the second law of thermodynamics, which states that entropy must increase in a closed system. Black holes contribute to entropy growth through Hawking radiation, defined by the Bekenstein-Hawking entropy formula:

$$S = \frac{kc^3 A}{4G\hbar}$$

where A is the area of the event horizon (Hawking 1974).

White holes, on the other hand, appear to violate this principle by ejecting ordered matter, effectively decreasing entropy. This contradiction suggests that if white holes exist, they must be

governed by principles beyond classical thermodynamics, likely within the framework of quantum gravity.

3. White Holes in Cosmology

3.1 The Big Bang as a White Hole

Some cosmologists propose that the Big Bang itself may have been a white hole event. This idea aligns with the notion that a white hole expels matter outward, much like the expansion observed in the early universe. The metric expansion of space in the standard ΛCDM model follows:

$$a(t) \propto t^{2/3(1+w)}$$

where a(t) is the scale factor and is the equation of state parameter (Schwarzschild 1916).

If the Big Bang was a white hole, it would mean our universe exists within the event horizon of a white hole expelling matter and energy from a higher-dimensional region. However, this remains speculative, as no clear link between white hole solutions and cosmic inflation has been established.

3.2 Primordial White Holes

It has been hypothesized that primordial white holes, formed in the early universe, may still exist as exotic remnants. These objects would have formed in high-density regions and could manifest as high-energy astrophysical phenomena.

3.3 Observational Challenges

Unlike black holes, which can be detected through gravitational effects and accretion disks, white holes would be extraordinarily difficult to observe. Their predicted repulsive nature suggests they would have no interacting matter nearby. Some researchers have speculated that certain fast radio bursts (FRBs) could be linked to transient white hole phenomena, but no definitive evidence has been found.

4. Quantum Gravity and White Holes

4.1 Loop Quantum Gravity Models

Loop quantum gravity (LQG) suggests that black holes might transition into white holes at the end of their evaporation process. This theory proposes that the singularity inside a black hole is replaced by a quantum bridge leading to a white hole, resolving the information paradox. The quantum bounce mechanism is modeled by:

$$H\Psi = 0$$

where is the H Hamiltonian constraint in LQG (Hawking 1974). This suggests that instead of complete evaporation, a black hole could transform into a white hole, gradually releasing stored information.

4.2 Information Paradox and White Holes

The black hole information paradox arises because black holes seem to destroy information, contradicting quantum mechanics. Some researchers propose that white holes serve as the mechanism through which information is eventually re-emitted into the universe, maintaining quantum coherence.

Materials and methods

Theoretical exploration and analysis were conducted through a detailed review of Einstein's field equations, specifically focusing on the Schwarzschild and Kerr metrics. A mathematical approach was used to discuss the time-reversed nature of white holes and their potential connections to wormholes. Observational methods and challenges were analyzed based on current astrophysical techniques and their applicability to the detection of white holes. References to existing models of loop quantum gravity and other theories relevant to quantum gravity were reviewed for potential insights into white holes.

Results

White holes were found to emerge as mathematical solutions to Einstein's field equations, specifically in the context of the Schwarzschild and Kerr metrics, with a time-reversed interpretation of the black hole solutions. The paper found that, in contrast to black holes, white holes expel matter and radiation, suggesting potential links to cosmological phenomena such as the Big Bang. Despite these theoretical connections, no empirical data currently supports the existence of white holes. Further theoretical exploration is needed to address their thermodynamic challenges, especially in relation to the second law of thermodynamics.

Discussion

The results suggest that white holes, while mathematically consistent, pose several theoretical challenges. The idea of white holes violating the second law of thermodynamics by expelling ordered matter presents a fundamental paradox. The possible connection between white holes and wormholes raises questions about the stability of such structures within classical general relativity and the need for exotic matter to stabilize them. In cosmology, the hypothesis that the Big Bang may have been a white hole event offers intriguing implications for our understanding of the universe's origins, though this idea remains speculative.

The unresolved black hole information paradox presents another avenue for research, with white holes potentially offering a resolution by re-emitting information into the universe. However, the lack of observational evidence remains a critical obstacle. Future research in quantum gravity and advancements in observational astrophysics may shed light on the plausibility of white holes.

Works Cited

Schwarzschild, K. (1916). "Über das Gravitationsfeld eines Massenpunktes nach der Einsteinschen Theorie." *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften*.

Hawking, S. W. (1974). "Black hole explosions?" *Nature*

Pascal's Triangle By Youqi Liu

Abstract

Pascal's Triangle is one of the most interesting results in mathematics. The value of an element found in row i and column j (where row and column numbers begin with zero) equals the number combinations that can be formed by choosing j objects from i objects. In spite of its apparent simplicity, the elements found in Pascal's Triangle are closely related to a large number of mathematical concepts, such as the coefficients of the terms in a binomial expansion. In addition, it is possible to derive key mathematical constants such as e, ϕ and π from the information found in Pascal's Triangle. Although Pascal's Triangle had previously been discovered by several mathematicians throughout history, Blaise Pascal was able to show many links between the triangle and other branches of mathematics. Over the past 350 years, mathematical research has continuously shown more interesting connections between Pascal's Triangle and other applications.

Introduction

Pascal's Triangle is a triangular array of non-negative integers; in a triangular array, the number of elements in a specified row cannot exceed the row number. For example, the fourth row can have a maximum of four elements.

In Pascal's Triangle, the rows are numbered from 0 to infinity; each element is indexed as (i, j) where i is the row number and j is the column number. The column elements range from 0 to the row number. For example, the first four rows are:

Row i = 0			1			
Row i = 1		1		1		
Row i = 2	1		2		1	
Row i = 3	1	3		3		1

The indices for these elements are:

Row i = 0			(0,0)			
Row i = 1		(1,0)		(1,1)		
Row i = 2	(2,0)		(2,1)		(2,2)	
Row i = 3	(3,0)	(3,1)		(3,2)		(3,3)
Column	j = 0	j = 1		j = 2		j = 3

The first row (i = 0) has one element; the second row (i = 1) has two elements; similarly the third row (i = 2) has three elements and the fourth row (i = 3) has four elements. The rest of the triangle is constructed in the same way.

The elements on each row of Pascal's Triangle are symmetrical from left to right. For example, the elements of row i = 3 are 1, 3, 3 1. The elements of row i = 7 are 1, 7, 21, 35, 35, 21, 7, 1. Another interesting property of Pascal's Triangle is known as "polarity." This means that by alternatively adding and subtracting the elements along a single row, the result is always zero. For example, in row i = 3, 1 + 3 – 3 + 1 = 0. In row i = 7, 1 + 7 – 21 + 35 – 35 + 21 – 7 + 1 = 0. Another interesting feature of Pascal's Triangle is that for all $n \geq 0$, row $2^n - 1$ has only odd elements. For example, if n = 2, then row $2^n - 1 = 2^2 - 1 = 3$ has the elements 1, 3, 3, 1. Similarly, if n = 3, then row $2^n - 1 = 2^3 - 1 = 7$ has the elements 1, 7, 21, 35, 35, 21, 7, 1.

There are two outer diagonals: the first is the set of the first elements in each row, and the second is the set of the last elements in each row. The elements on the outer diagonals are always equal to one.

0				1			
1			1		1		
2		1		2		1	
3	1		3		3		1

0				1			
1			1		1		
2		1		2		1	
3	1		3		3		1

Fig 1: Outer diagonals

The remaining elements are computed as the sum of the two adjacent elements in the previous row:

$$(i, j) = (i - 1, j - 1) + (i - 1, j) \qquad (1)$$

For example, the element (3,1) is the sum of the elements (2,0) and (2,1) which is 1 + 2 = 3.

Each element in Pascal's triangle equals the number of combinations that can be formed by choosing j objects from i objects.

$$_iC_j = \frac{i!}{j!(i-j)!} \qquad (2)$$

For example, the element (3,1) equals:

$$_3C_1 = \frac{3!}{1!(3-1)!} = \frac{3!}{1!2!} = \frac{6}{(1)(2)} = 3 \qquad (3)$$

These elements are known as "binomial coefficients" since they are the coefficients of a binomial expansion of $(x + y)^i$ where i is the row number. For example, the coefficients in row i = 3 are:

1, 3, 3, 1

These are the coefficients of the binomial expansion $(x + y)^3 = x^3 + 3x^2y + 3xy^2 + y^3$.

Pascal's Triangle contains a huge number of patterns that can be useful for many applications in the fields of combinatorics, algebra, probability theory and number theory.

Historical Background

Blaise Pascal (1623 – 1662) was a French philosopher, theologian, scientist, inventor and mathematician. Pascal published some influential papers while still in his teens, and through his correspondence with Rene Descartes and Pierre de Fermat influenced the development of probability theory as well as economics. He also invented a mechanical device designed to do basic calculations. In later life, Pascal dedicated more of his time to theological issues.

In 1654, Pascal's friend the Chevalier de Mere encouraged him to research the field of probability theory. This inspired Pascal to collaborate with Pierre de Fermat, leading to the development of many important results in this field. Pascal's treatise on the Arithmetical Triangle (now known as Pascal's Triangle) was written in 1654 but not published until 1665 after Pascal had died.

Pascal's Triangle was not a completely new idea, as several other authors throughout history had written about Arithmetic Triangles. The Chinese mathematician Jia Xian wrote about a similar triangle in the 11th century, followed by another Chinese mathematician Yang Hui in the 13th century. In Persia, the 11th century mathematician Omar Khayyam used a similar concept to describe binomial expansions. In India, a triangle closely related to Pascal's Triangle was developed by Meru Prastara around 300 BCE to describe the pattern of syllables in poetry. Pascal was able to formalize these ideas and develop them further, showing the relationship between Pascal's Triangle and the field of combinatorics.

Interesting Patterns in Pascal's Triangle

One of the most fascinating aspects of Pascal's Triangle is the large number of patterns that can be found in the different diagonals of the triangle. Although the elements of Pascal's triangle represent combinations, there are many other potential applications of these numbers. These include:

- prime numbers
- triangular numbers
- powers of 2
- natural numbers
- tetrahedral numbers

- pentatope numbers
- Fibonacci numbers

Prime Numbers

An interesting result found in Pascal's Triangle is that for the row numbers that are prime numbers, excluding the first and last elements of the row (which must be ones) all other row elements are divisible by the row number.
For example, the row i = 5 in Pascal's Triangle is:

1 5 10 10 5 1

Each element in this row (aside from the first and last) is divisible by 5.
As another example, the row i = 7 in Pascal's triangle is:

1 7 21 35 35 21 7 1

Each element in this row (aside from the first and last) is divisible by 7.
This holds whenever a row number is a prime number; it is not true when the row number is a composite number.

Triangular Numbers

Triangular numbers are the numbers of objects that can be formed into an equilateral triangle. For example:

$T_1 = 1, T_2 = 3, T_3 = 6, T_4 = 10, T_5 = 15, T_6 = 21, T_7 = 28$

In general, these can be computed as: $T_n = {}_{n+1}C_2 = n(n+1)/2$

Based on this definition, the triangular numbers can be found in the third diagonal of Pascal's Triangle:

											Triangular Numbers									
0												1								
1											1		1							
2										1		2		1						
3									1		3		3		1					
4								1		4		6		4		1				
5							1		5		10		10		5		1			
6						1		6		15		20		15		6		1		
7					1		7		21		35		35		21		7		1	
8				1		8		28		56		70		56		28		8		1

Fig 2: Diagonal of Pascal's Triangle

For example, the fifth triangular number is: $T_5 = n(n+1)/2 = 5(6)/2 = 30/2 = 15$
The element (6,2) contains 15; in general, this result will be found in element (n + 1, 2).

Powers of Two

Each row generates a power of two as the sum of the elements on that row. This is shown in the following diagram:

0					1					$1 = 2^0 = 1$			
1				1		1				$1+1 = 2^1 = 2$			
2			1		2		1			$1+2+1 = 2^2 = 4$			
3		1		3		3		1		$1+3+3+1 = 2^3 = 8$			

Fig 3: Power-of-Two Summation Triangle

For example, in the row i = 2 the elements are 1, 2, 1; their sum is 4, which is 2^2. This is true for all rows in Pascal's Triangle.

Natural Numbers

The natural numbers are the positive integers: N = {1, 2, 3, 4, ….}. These can be found on the second diagonal of Pascal's Triangle:

								Natural Numbers									
i																	
0									1								
1								1		1							
2							1		2		1						
3						1		3		3		1					
4					1		4		6		4		1				
5				1		5		10		10		5		1			
6			1		6		15		20		15		6		1		
7		1		7		21		35		35		21		7		1	
8	1		8		28		56		70		56		28		8		1

Fig 4: The natural numbers

Due to the symmetry of Pascal's Triangle, these numbers can also be found as:

								Natural Numbers									
i																	
0									1								
1								1		1							
2							1		2		1						
3						1		3		3		1					
4					1		4		6		4		1				
5				1		5		10		10		5		1			
6			1		6		15		20		15		6		1		
7		1		7		21		35		35		21		7		1	
8	1		8		28		56		70		56		28		8		1

Fig 5: The natural numbers

Tetrahedral Numbers

The tetrahedral numbers are the numbers that can be used to form a pyramid with a triangular base (known as a tetrahedron.) The first six tetrahedral numbers are 1, 4, 10, 20, 35, 56, …. These are computed as:

$$T_n = \sum_{k=1}^{n} \left(\frac{k(k+1)}{2} \right)$$

For example, T_3 =

$$T_3 = \left(\frac{1(1+1)}{2} \right) + \left(\frac{2(2+1)}{2} \right) + \left(\frac{3(3+1)}{2} \right)$$

$$T_3 = \left(\frac{2}{2} \right) + \left(\frac{6}{2} \right) + \left(\frac{12}{2} \right) = 1+3+6 = 10$$

These are shown in the fourth diagonal of Pascal's Triangle:

								Tetrahedral Numbers									
i																	
0									1								
1								1		1							
2							1		2		1						
3						1		3		3		1					
4					1		4		6		4		1				
5				1		5		10		10		5		1			
6			1		6		15		20		15		6		1		
7		1		7		21		35		35		21		7		1	
8	1		8		28		56		70		56		28		8		1

Fig 6: The tetrahedral Numbers

Pentatope Numbers

The first five pentatope numbers are: 1, 5, 15, 35, 70. These are computed as:

$$P_n = \frac{n(n+1)(n+2)(n+3)}{24}$$

For example, the third pentatope number is computed as:

$$P_3 = \frac{3(3+1)(3+2)(3+3)}{24} = \frac{360}{24} = 15$$

The pentatope numbers are found in the fifth diagonal of Pascal's Triangle:

								Pentatope Numbers									
i																	
0									1								
1								1		1							
2							1		2		1						
3						1		3		3		1					
4					1		4		6		4		1				
5				1		5		10		10		5		1			
6			1		6		15		20		15		6		1		
7		1		7		21		35		35		21		7		1	
8	1		8		28		56		70		56		28		8		1

Fig 7: The pentatope numbers

Fibonacci Numbers

If Pascal's Triangle is expressed in left-justified form, the terms of the Fibonacci sequence can be obtained as sums of the resulting diagonals.

The Fibonacci Sequence is a recursive sequence in which each term (aside from the first two) is the sum of the previous two terms:

1, 1, 2, 3, 5, 8, 13, 21, 34, …..

						Fibonacci Numbers					
0			1								
1			1	1							
2			1	2	1						
3			1	3	3	1					
4			1	4	6	4	1				
5			1	5	10	10	5	1			
6			1	6	15	20	15	6	1		
7			1	7	21	35	35	21	7	1	
8			1	8	28	56	70	56	28	8	1

Fig 8: The fibonacci sequence

For example, the term 13 is the sum of:

						Fibonacci Numbers					
0			1								
1			1	1							
2			1	2	1						
3			1	3	3	1					
4			1	4	6	4	1				
5			1	5	10	10	5	1			
6			1	6	15	20	15	6	1		
7			1	7	21	35	35	21	7	1	
8			1	8	28	56	70	56	28	8	1

Fig 9: The fibonacci numbers

Deriving Constants from Pascal's Triangle

Several key constants found in mathematics, such as e, ϕ and π, can be derived using Pascal's Triangle.

Deriving e

e (also known as Euler's constant) is approximately equal to 2.71828…. e appears in many applications, such as finance where it is used to compute the future value of a sum where

the interest rate is compounded continuously. It can be used to model exponential growth and exponential decay in scientific applications.

The value of e can be derived from Pascal's triangle as follows. The first step is to compute the products of the elements along each row. For example:

row i = 0 1
row i = 1 1 * 1 = 1
row i = 2 1 * 2 * 1 = 2
row i = 3 1 * 3 * 3 * 1 = 9
row i = 4 1 * 4 * 6 * 4 * 1 = 96
row i = 5 1 * 5 * 10 * 10* 5 * 1 = 2500

The next step is to compute the ratios of consecutive row products. These results are not interesting by themselves, but by computing the ratios of these ratios, an interesting pattern emerges:

Row (i)	Product	Ratios	Ratios of Ratios
0	1	1.00	2.00
1	1	2.00	2.25
2	2	4.50	2.37
3	9	10.67	2.44
4	96	26.04	2.49
5	2500	64.80	2.52
6	162000	163.40	2.55
7	26471025	416.10	****
8	11014635520	****	****

Table 1: Ratio Cascade Pattern

Each ratio is obtained as: ratio i = row i+1 product / row i product
For example, ratio 3 equals row 4 product divided by row 3 product which is 96/9 = 10.67.
The ratios of ratios are obtained as: ratio of ratio i = ratio i + 1 / ratio i
The ratio of ratio 3 is obtained as: ratio 4 / ratio 3 = 26.04/10.67 = 2.44

Looking at the ratios of ratios, the numbers get progressively closer to e (2.71828….) as i increases. In the limit, as i approaches infinity, the ratio of ratios approaches e.

Deriving ϕ

ϕ is a constant that is equal to $\frac{1+\sqrt{5}}{2} = 1.618033.....$ It is often referred to as the "Golden Ratio" and appears in a surprisingly large number of different applications, such as math, biology, art, architecture, botany, etc. ϕ can be derived in many different ways; for

example, it is the limit of the ratio of consecutive terms in the Fibonacci Sequence, which is defined as:

$$F_n = F_{n-1} + F_{n-2} \quad \text{where } F_1 = 1 \text{ and } F_2 = 1$$

The following table shows the consecutive ratios of the terms of the Fibonacci Sequence:

n	Fn	Fn+1 / Fn
1	1	1.0000
2	1	2.0000
3	2	1.5000
4	3	1.6667
5	5	1.6000
6	8	1.6250
7	13	1.6154
8	21	1.6190
9	34	1.6176
10	55	1.6182

Table 2: Fibonacci Consecutive Ratio Table

This shows that as n increases, the ratio of consecutive Fibonacci terms F_{n+1}/F_n rapidly approaches $\phi = 1.618033\ldots$. These results can be derived from Pascal's Triangle as follows.

By rewriting Pascal's triangle with diagonals converted into columns, the terms of the Fibonacci Sequence are found from drawing a line through consecutive terms on a diagonal and adding. For example, the following example shows how the terms 13 and 34 are found:

						Fibonacci Numbers						
0			1									
1			1	1								
2			1	2	1		**13**					
3			1	3	3	1		**34**				
4			1	4	6	4	1					
5			1	5	10	10	5	1				
6			1	6	15	20	15	6	1			
7			1	7	21	35	35	21	7	1		
8			1	8	28	56	70	56	28	8	1	
9			1	9	36	84	126	126	84	36	9	1
10			1	10	45	120	210	252	210	120	45	10	1

Fig 10: Fibonacci discovery in transformed pascal's triangle

By determining each term in the Fibonacci Sequence and computing the ratios of consecutive terms, the value of ϕ is obtained.

Deriving π

π is one of the most well-known constants in mathematics, appearing in many formulas in Geometry such as the area of a circle = A = πr^2.

π can be obtained by starting with 2 and then alternatively adding and subtracting the ratio of 1 over consecutive triangular numbers. The triangular numbers are found on the third diagonal of Pascal's Triangle:

								Triangular Numbers									
i																	
0									1								
1								1		1							
2							1		2		1						
3						1		3		3		1					
4					1		4		6		4		1				
5				1		5		10		10		5		1			
6			1		6		15		20		15		6		1		
7		1		7		21		35		35		21		7		1	
8	1		8		28		56		70		56		28		8		1

Fig 11: Deriving π Using Triangular Numbers in Pascal's Triangle

π is obtained as follows:

$$\pi = 2 + \left(\frac{1}{1} + \frac{1}{3} - \frac{1}{6} + \frac{1}{10} - \frac{1}{15} + \frac{1}{21} - \frac{1}{28} .... \right)$$

Using the first seven triangular numbers gives 3.345238…

By using a progressively larger number of Triangular numbers, this expression will converge to the value of π in the limit.

Works Cited

"Pascal's Triangle", Wikipedia. https://en.wikipedia.org/wiki/Pascal%27s_triangle

"Pascal's Triangle", Mathisfun.com.
https://www.mathsisfun.com/pascals-triangle.html

"Applications of Pascal's Triangle" Neurochispas.com.
https://en.neurochispas.com/algebra/applications-of-pascals-triangle/

History of Pascal's Triangle

"Pascal's Triangle", Historymath.com. https://www.historymath.com/pascals-triangle/

Biography of Blaise Pascal

"Blaise Pascal", Wikipedia. https://en.wikipedia.org/wiki/Blaise_Pascal

Patterns in Pascal's Triangle

Triangular Numbers

"Triangular Numbers", Wikipedia.com. https://en.wikipedia.org/wiki/Triangular_number

"Pascal's Triangle", GoldenNumber.net.
https://www.goldennumber.net/pascals-triangle/

"The Mathematical Secrets of Pascal's Triangle", YouTube.
https://video.search.yahoo.com/search/video?fr=mcafee&p=applications+of+pascal%27s+triangle&type=E210US0G0#id=1&vid=4ae21f3191cdf809bb53193eaa867b2e&action=click

The Rise of Asset Bubbles and Policy Responses By Wentao Zhang

Abstract

An asset bubble arises when market participants begin buying up assets for speculative purposes, driving up their prices to levels that are inconsistent with market fundamentals. Although each bubble is unique, there are many common underlying factors that lead to the creation of these bubbles. Often people are attracted by assets that promise huge future profits and begin to believe that there is little or no risk in buying these assets. At some point, market participants begin to recognize that these assets are extremely overvalued, leading to a market selloff and a collapse of the asset price. There have been several classical examples of this throughout history, such as Tulipmania in the Netherlands in the 17th century; more recent examples include the dot.com bubble of the late 1990's and the Housing Crisis of 2007-8. The Federal Reserve has publicly stated that they are not in a position to detect or prevent the formation of asset bubbles. As a result, asset bubbles are likely to arise in the future since human nature does not change.

Introduction

A financial bubble is a situation in which the price of a financial asset experiences a sudden, dramatic increase in value. This sudden increase is heavily driven by market psychology rather than economic fundamentals. At some point, when market sentiment changes, the price of the asset crashes back to more normal levels.

There have been many examples of this throughout history, the most well-known being Tulipmania in the Netherlands in the 17th Century. During the period from November 1636 to February 1637, the price of certain tulip bulbs increased dramatically to the point where a single bulb could be more expensive than a house. In February 1637 the market collapsed and by May 1637 tulip prices had fallen back to their original levels, causing massive financial losses for the investors.

Another well-known financial bubble was the South Sea bubble, which took place in the early 18th century. The South Sea Company was founded by the U.K. Parliament in 1711 and was granted a trade monopoly for North America. The company paid investors an annual dividend of 6%, which was based on expected future profits from North American trade. King George I became governor of the company in 1718, which increased confidence in the company, leading to more speculative buying of the stock. In 1720 Parliament sold the national debt of 32 million pounds to the South Sea company for 7.5 million pounds. The company was supposed to keep paying interest on the debt from its future profits. In August 1720 the price of the stock reached 1,000 pounds per share (equivalent to 233,076 pounds in 2024 terms). Unfortunately, the South Sea company did not have nearly enough earnings to cover its commitments, and the price fell off to 124 pounds per share by December 1720.

A more recent example is the rapid runup of housing prices between 2005-2007 when easy credit made it possible for more people to buy homes even when their financial situation

was not strong enough to qualify for a mortgage based on historical standards. The result was a huge surge of housing prices which led to a rapid collapse in 2007-8, causing the bankruptcy of many financial institutions.

What is an Asset Bubble?

Asset bubbles arise when the demand for a product increases dramatically, causing the price to rise rapidly even without any improvement in the fundamental factors that determine its value. Asset bubbles can arise for many reasons. For example, the Housing Crisis of 2007-8 was caused by several factors such as unusually low interest rates, extremely loose lending standards by banks and a pervasive sense by investors that housing prices are highly unlikely to fall. In addition, banks introduced many complex mortgage products that were designed to enable people to buy houses based on the expectation of future price increases. For example, Adjustable Rate Mortgages (ARMs) were designed to offer extremely low mortgage payments for several years, followed by rapid increases in the mortgage payments. This encouraged people with relatively low incomes to buy houses based on the expectation of rapid price appreciation and rising future incomes. These factors encouraged many people to buy homes who normally would not have had the appropriate credit rating and/or income necessary to buy a home. The result was a huge increase in the demand for housing for speculative purposes without any corresponding increase in the size of the population or the supply of housing. The bubble burst when a large percentage of these homeowners were unable to make their mortgage payments in a timely manner. This led to huge losses among banks which restricted lending and led to an ugly recession in the United States.

At the same time, the introduction of credit derivatives convinced investors that they were protected from default risk, which turned out to be a false assumption. The major credit rating agencies such as Standard and Poor's made this situation worse by giving "investment grade" ratings to credit derivatives that were not justified. Another major factor in the housing crisis was the actions of the government mortgage agencies Fannie Mae and Freddie Mac; these organizations used subprime mortgages in the mortgage pools used as collateral for securities such as mortgage-backed securities, thereby dramatically increasing the risk of these securities. Many investors lost enormous sums of money from these securities which had historically been extremely safe.

The "dot-com" bubble occurred in the late 1990's when the prices of internet stocks rose dramatically, causing the NASDAQ to rise by 800% from 1995 to 2000 before falling by 78% by October 2022. The index reached a peak of 5132.52 on March 10, 2000 and fell to 1,100 by October 2022. The NASDAQ (National Association of Securities Dealers Automated Quotations) consists primarily of high technology stocks so its price was dramatically impacted by the surge of demand for these stocks. Many of these companies started up in the late 1990's following the availability of the World Wide Web (WWW) which brought internet access to individuals. The stocks of high-technology companies such as Apple, Microsoft, Amazon.com, and Google increased dramatically as investors sought to take advantage of the rapidly

advancing technology and the expected future profitability of these companies. The prices increased to levels that were not consistent with the underlying economic fundamentals of these companies so that a bubble formed based on what Fed Chairman Alan Greenspan called "irrational exuberance." Following the December 1996 speech in which this phrase was used, the stocks of the high-tech companies began to fall dramatically, leading to a deep recession in 2001. This asset bubble had much in common with Tulipmania as investors paid exorbitant prices for these stocks based on their hopes of dramatic future profits. As investors began to realize that the stocks were overvalued, a surge of selling led to a collapse of these stock prices.

Dramatic changes in the supply of an asset can also lead to asset bubbles. Some recent examples are the rapid increase in oil prices from 1999 to 2008, when reduced production by Saudi Arabia combined with continuously growing demand around the world, especially in China, caused the price of crude oil to rise from less than $25 per barrel to over $200 per barrel from early 1999 to the summer of 2008. The price fell back below $25 per barrel during the out break of COVID in early 2020.

The crypto currency Bitcoin also experienced many surges in price during its brief history. Bitcoin was first introduced in 2009, with one of the first transactions being the purchase of two pizzas from Papa John's for 10,000 bitcoins. During this time, bitcoin's price has ranged from about $250 to over $100,000 in 2025. There are several reasons to doubt that bitcoin's price history represents an asset bubble. For example, each time the price drops significantly, it eventually rebounds to set a new high price. There is a well-established demand for bitcoins, and their supply is limited to 21 million coins which helps ensure that their price can't completely collapse. Many institutional investors hold bitcoin due to its potential price appreciation, which wouldn't happen if they were concerned about bubbles. In countries with unstable political or banking systems bitcoins are in strong demand as protection against a collapse of the local currency.

In the early 1990's the Japanese economy was severely damaged by the bursting of a real estate bubble. Japan's economy entered a recession in 1986 as the strong yen reduced Japan's trade surplus. In response the Bank of Japan lowered interest rates to stimulate the economy. Speculators used these low interest rates to buy up real estate, causing prices to surge to unheard of levels. At one point the land surrounding the Imperial Palace in Tokyo was valued at a higher price than all the real estate in California. The bubble burst in 1991, causing a massive recession. Real estate prices did not recover to their pre-bubble levels until 2018.

Common Features of Asset Bubbles

Asset bubbles can arise if there is a dramatic increase in the demand for an asset, such as tulip bulbs during Tulip Mania or housing during the U.S. housing crisis of 2007-2008 or high tech stocks during the dot.com bubble of the late 1990's. In these cases, the surge in demand far outstrips the growth in supply of the asset, causing the price to rise to levels that are inconsistent with fundamental factors. In these cases, there comes a point where some of the investors

attempt to sell their holdings before prices fall; once this process begins, there is a dramatic sell-off and the price of the asset collapses.

Asset bubbles can also arise due to a sudden reduction in the supply of an asset, such as the oil bubble of the late 1990's – early 2000's. In these cases, the price can continue to rise to new record levels and will only fall back to more normal levels when there is an increase in supply or a reduction in demand for the asset or a discovery of alternative supplies.

Another common feature of bubbles is that people are typically unaware that a bubble is forming since it is difficult to tell if an asset is overpriced. In particular, when new technology is introduced to the market, normal valuations are more difficult to carry out due to the increase in uncertainty over the future value of assets related to the new technology.

The Federal Reserve and Bubbles

The Federal Reserve has publicly discussed the issue of asset bubbles and whether it is possible for the Fed to take any steps to prevent bubbles from forming. The President of the Minneapolis Federal Reserve Bank, Neel Kashkari, explained in his 2017 article "Monetary Policy and Bubbles" that "it is really hard to spot bubbles with any confidence before they burst." [1] He goes on to point out that managing bubbles is not part of the Fed's mandate and that the cost of policy mistakes is extremely high, so the Fed proceeds cautiously when a bubble appears to be forming. In addition, he states that stopping bubbles is difficult because "future generations are exceptionally good at repeating past mistakes." He also explains why bubbles form in the first place as "human societies are prone to mass delusion and bubbles."

According to the author, the Fed's mandate includes the conduct of monetary policy to ensure stable prices and maximum employment, along with financial stability. In addition, the Fed is responsible for regulating the banking system. Managing asset bubbles is not part of the Fed's mandate. Using interest rates to prevent bubbles from forming could have enormous adverse consequences such as causing a recession without necessarily addressing the root causes of the bubble. Further, the Fed's policy tools are not well-suited for preventing bubbles. In addition, these policy tools affect the entire economy and could be counter-productive if not used correctly. The Fed's policy tools are better suited to deal with the fallout of a bubble that has burst than to prevent bubbles from forming in the first place. For example, after the Housing Crisis of 2007-2008 the Fed engaged in a process known as "quantitative easing" in which it provided liquidity to the banking system by buying "non-performing" assets from the banks such as Credit Default Swaps and Collateralized Debt Obligations. This enabled the banks to continue doing business and prevented a catastrophic failure of the entire U.S. financial system.

The author also makes the case that even when attempts are made to prevent a bubble from forming, the efforts may not be successful. For example, in October 2010 the regulatory authorities in Sweden attempted to slow the rapid growth in housing prices by increasing the loan-to-value ratio required for new mortgages. This failed to slow the rapid growth of housing prices. In Vancouver, British Columbia the local authorities passed a 15% tax on foreign

[1] "Monetary Policy and Bubbles", Neel Kashkai. May 17, 2017. Federal Reserve Bank of Minneapolis.

purchasers of housing in August 2016 to slow the rapid price appreciation of local housing. This step was unsuccessful in slowing the rise of housing prices. These examples show how difficult it would be for any government agency to slow the rise of prices for housing or any other financial assets.

The author also makes the point that for many asset bubbles, the only people who lose out are the direct participants in the bubble, which was the case with Tulipmania. In that case, it would not be good public policy to interfere with the economy to prevent the rise of bubbles as the costs could greatly exceed the benefits. The impact on the economy when a bubble finally bursts depends heavily on the use of leverage; i.e., how much borrowed money was used to buy financial assets. In particular, the housing crisis of 2007-2008 was particularly severe since many houses were purchased with very low or non-existent down payments and were purchased by people with sub-prime credit ratings.

Overall, the author does not think that it is appropriate for the Fed to attempt to prevent bubbles from forming due to the difficulty of identifying them and the inadequacy of its policy tools to stop them.

Conclusion

There have been many asset bubbles throughout history, such as Tulipmania, the South Sea Bubble, the Housing Crisis of 2007-8, etc. Each of these has many features in common: people pay exorbitant prices for assets that exceed their fundamental value in hopes of earning speculative profits. Once a critical point has been reached, a large number of investors attempt to sell their assets and realize their profits at the same time, causing a collapse of the bubble. The Federal Reserve has been criticized for not attempting to prevent bubbles from forming, but this is outside of its mandate and they do not have the appropriate policy tools to make this happen.

In the future, assets bubbles are likely to arise again, as human nature doesn't change. Although each case appears to be "different" from past cases, the underlying causes are always very similar. At the moment, reminiscent of the dot.com bubble, investors are becoming excited about the rise of Artificial Intelligence and the stocks of the companies that have invested heavily in this area. For example, stocks such as:

- AppLovin
- NVIDIA
- Microsoft
- Apple
- Tesla
- Amazon
- Meta
- Google
- OpenAI

have all experienced significant growth in the past few years, partially due to their Artificial Intelligence activities. For example, the market capitalization of NVIDIA rose from just over $100 billion in late 2019 to over $3 trillion at the beginning of 2025. Over this same period Apple's market capitalization rose from just under $1 trillion to more than $3.5 trillion. Alphabet's market capitalization during this time rose from just under $1 trillion to over $2.5 trillion. During this same time frame the NASDAQ has experienced extremely rapid growth rising from about 7,000 in May 2020 to a record high of over 20,000 in January 2025. Much of this growth has been fueled by rapidly advancing technology, particularly in the field of Artificial Intelligence.

It is an interesting question to consider whether a bubble is developing among high-tech stocks or if this is simply a reflection of the advancing technology, which has enormous implications for many different industries. As with past bubbles, it is difficult to be sure if this represents more normal rates of return or if this is another bubble developing.

Works Cited

What is an Economic Bubble?

"What is an Economic Bubble and How Does it Work, With Examples?" Will Kenton, April 3, 2022. Investopedia.com. https://www.investopedia.com/terms/b/bubble.asp#:~:text=A%20bubble%20is%20an%20economic%20cycle%20that%20is,to%20as%20a%20%22crash%22%20or%20a%20%22bubble%20burst.%22

Historical Bubbles

Tulipmania

"Tulipmania", Wikipedia.com. https://en.wikipedia.org/wiki/Tulip_mania
https://www.history.com/news/tulip-mania-financial-crash-holland

"The Real Story Behind the 17th Century 'Tulip Mania' Financial Crash", Dave Roos. August 4, 2023. https://www.history.com/news/tulip-mania-financial-crash-holland

"Tulipmania (1637)", Economics and History.substack.com. August 28, 2024. https://economicsandhistory.substack.com/p/tulip-mania-1637?subscribe_prompt=free

YouTube videos: https://www.youtube.com/results?search_query=the+history+channel+tulipmania)

South Sea Bubble

"The South Sea Bubble," Terry Stewart, History-UK.com. https://www.historic-uk.com/HistoryUK/HistoryofEngland/South-Sea-Bubble/

Recent Bubbles

Dot Com Bubble

"Dot-com bubble", Wikipedia.com. https://en.wikipedia.org/wiki/Dot-com_bubble

Housing Bubble Late 2000's

"2000s United States Housing Bubble", Wikipedia.com. https://en.wikipedia.org/wiki/2000s_United_States_housing_bubble

Bitcoin

"Cryptocurrency Bubble", Wikipedia.com. https://en.wikipedia.org/wiki/Cryptocurrency_bubble#:~:text=A%20cryptocurrency%20bubble%20is%20a%20phenomenon%20where%20the,bubbles%20on%20a%20boom%20to%20bust%20cycle.%20%5B1%5D%5B2%5D

Bitcoin Price History

https://finance.yahoo.com/quote/BTC-USD/

"Bitcoin's Bubbly Behaviors: Does it Resemble Other Financial Bubbles of the Past?", Sergio Alonso, et.al. June 3, 2024. Nature.com. https://www.nature.com/articles/s41599-024-03220-0

Federal Reserve Policy

"Irrational Exuberance"

"Irrational Exuberance", Wikipedia.com.
https://en.wikipedia.org/wiki/Irrational_exuberance
The Fed and Asset Bubbles
"Bubbles and Stagnation", Ines Xavier, May 2022. Federal Reserve Board of Governors. https://www.federalreserve.gov/econres/feds/bubbles-and-stagnation.htm
"How Should We Respond to Asset Price Bubbles?", Frederic Mishkin, May 15, 2008. Federal Reserve Board of Governors. https://www.federalreserve.gov/newsevents/speech/mishkin20080515a.htm
"Monetary Policy and Bubbles", Neel Kashkari, Federal Reserve Bank of Minneapolis. May 17, 2017. https://www.minneapolisfed.org/article/2017/monetary-policy-and-bubbles
Causes of Asset Bubbles
"What Causes Asset Bubbles?", Amy Fontinelle, November 25, 2024. https://www.investopedia.com/financial-edge/0911/what-causes-bubbles.aspx
"Five Stages of a Bubble", Troy Segal, November 2, 2024. https://www.investopedia.com/articles/stocks/10/5-steps-of-a-bubble.aspj
"Asset Bubbles: Causes and Trends," Kimberly Amadeo, November 22, 2021. The Balance. https://www.thebalancemoney.com/asset-bubble-causes-examples-and-how-to-protect-yourself-3305908
"Is Bitcoin a Bubble? Sure, if Bubbles Could Last for Decades", Amir Tabch. November 27, 2024. Middle East Economy. https://economymiddleeast.com/news/is-bitcoin-a-bubble-sure-if-bubbles-could-last-for-decades/
"Tulip Mania: The Flowers That Cost More Than Houses", Alastair Sooke, May 3, 2016. BBC. https://www.bbc.com/culture/article/20160419-tulip-mania-the-flowers-that-cost-more-than-houses
"At One Point, Amazon Lost More Than 90% of its Value. But Long-Term Investors Still Got Rich.", Andrew Davis, December 18, 2018. CNBC.com. https://www.cnbc.com/2018/12/18/dotcom-bubble-amazon-stock-lost-more-than-90percent-long-term-investors-still-got-rich.html
"Asset Bubbles Throughout History: The 5 Biggest", Elvis Pichardo, April 19, 2022. https://www.investopedia.com/articles/personal-finance/062315/five-largest-asset-bubbles-history.asp

The VIew of Artificial Intelligence on Causes of Heart Disease
By Akash Sharma

Introduction

Around 17.9 million people die from heart disease every year (WHO,n.d.). This has led to a growing worldwide need to find out what is causing heart disease. So, to contribute to that cause, I created an artificial intelligence (AI) model that was trained on an excerpt of data provided by the Behavioral Risk Factor Surveillance System (BRFSS) (Teboul). Based on the AI model, I found that general health was one of the leading causes of heart disease.

Literature Review: Heart Disease

Heart disease is a disease which affects the heart by changing the structure of the heart or by changing the function of it. Heart disease is a very general term with many branches within it. One branch, coronary artery disease, the most common one, affects how blood flows to the heart. Specifically, it is the decrease in blood flow to the heart. This can happen due to a build-up of cholesterol, fat, or other substances in the arteries and veins of the heart, also called atherosclerosis (Heart and Stroke Foundation of Canada).

Another factor that affects coronary heart disease is the amount of low-density lipoprotein or LDL cholesterol. This type of cholesterol is the cholesterol that sticks to the arteries and veins of the circulatory system, causing blockages and leading to coronary artery disease. A combating force to LDL cholesterol is High-Density Lipoprotein or (HDL) cholesterol because it carries cholesterol through the body and to the liver where it is removed.

Cholesterol functions in the human cell mainly through cholesterol rafts, which are areas with high cholesterol concentrations that allow the cell to interact with other cells. Cholesterol in the cell is also divided into detergent-sensitive and detergent-resistant parts. Detergent-sensitive cholesterol can be broken down with mild detergents, while detergent-resistant cholesterol is not broken down with those mild detergents. An experiment was done on rats to determine the effects of cholesterol on various arteries in the body (Schneider et al.). Cholesterol used actively within a cell is known as esterified cholesterol and cholesterol combined with a fatty acid for storage in the body is known as unesterified cholesterol. Detergent-sensitive, detergent-resistant, esterified, and unesterified cholesterol levels were measured in rats with a high-cholesterol diet to find how different arteries accumulate cholesterol. The study found that HDL cholesterol does not increase with a high-cholesterol diet whereas LDL cholesterol does increase. They also found that the liver-to-body-mass ratio increased with an increase in the amount of cholesterol in the liver.

However, there are ways to combat the effects of cholesterol on the body. One innovative and upcoming way is Reverse Cholesterol Transport (RCT), which involves the process of reversing LDL transport into arteries and instead transporting cholesterol into the liver and then out with fecal excretion (Rosenson et al.).

Another type of heart disease is rheumatic heart disease. Rheumatic heart disease is caused by damage to the blood's passageways into the heart caused by rheumatic fever. Streptococcus pyogenes cause rheumatic fever. The diagnostic criteria for heart disease require either two major, or one major and two minor criteria to be considered as rheumatic fever. Additional evidence is also required like a pharyngeal swab culture that returns positive for Group A β hemolytic Streptococcus (GABHS). The criteria for acute rheumatic fever is carditis, which is typically just valvulitis, polyarthritis, swelling, and pain in joints, the location of which changes as the disease continues; Sydenham Corea, which is the quick, involuntary movement of the trunk or hands and feet; erythema marginatum which is a snake-like rash also found on the trunk or hands and feet; and subcutaneous nodules, which are bumps found on arms and legs. Some additional side effects include behavioral changes and instability of emotions. The minor criteria for the disease are fever, specifically at least 39℃, first-degree heart block, arthralgias, and increased acute-phase reactants. First-degree heart block delays the brain's electrical signals to the heart. Acute-phase reactants are markers of inflammation in the body, typically increasing when a person has rheumatic heart disease (Seckeler and Hoke).

Heart disease in the early 20th century was quite low, being uncommon to get. However, by the mid-20th century, it became the leading cause of death for Americans which, into the late 20th century, had continued to reduce. There are many theories as to why the rates of heart disease have lowered over the past few decades, with one related to the decrease of coronary atherosclerosis, a factor of which is smoking. There have also been many fewer people who have been accepted into hospitals for acute myocardial infarction, which could be due to young people being exposed to fewer risk factors. Due to various new therapy techniques, many people are surviving acute myocardial infarction and various types of heart disease (Dalen et. al., 2014).

Literature Review: Random Forest Classification

A random forest classifier is made up of many different decision trees, which can be changed to vary the complexity of the model. A decision tree works by trying to separate data into different boxes to try and find the classification of each item or an "answer". Once the tree has found an answer, the random forest averages all of the answers given by each decision tree in the forest. It then outputs this answer, and that is the random forest classifier's prediction. To make a random classifier, first, the random forest classifier must go through a training phase. The training phase is essentially the model trying to minimize the loss function, which is the difference between a model's predicted value and the correct value. It does this by gradient descent, where the model will tweak itself so that it will find the parameters that result in the lowest value for the loss function. After this, one inputs a test set to "test" the random forest classifier and see how well it performs on unseen data (data that it was not trained on). For example, say someone wants to figure out who will be the MVP for a season in football, they would first need to input past years' data to train the random forest classifier. Then, once they have trained the classifier, and if they have a hopefully high enough accuracy, they can input this year's data and have an answer.

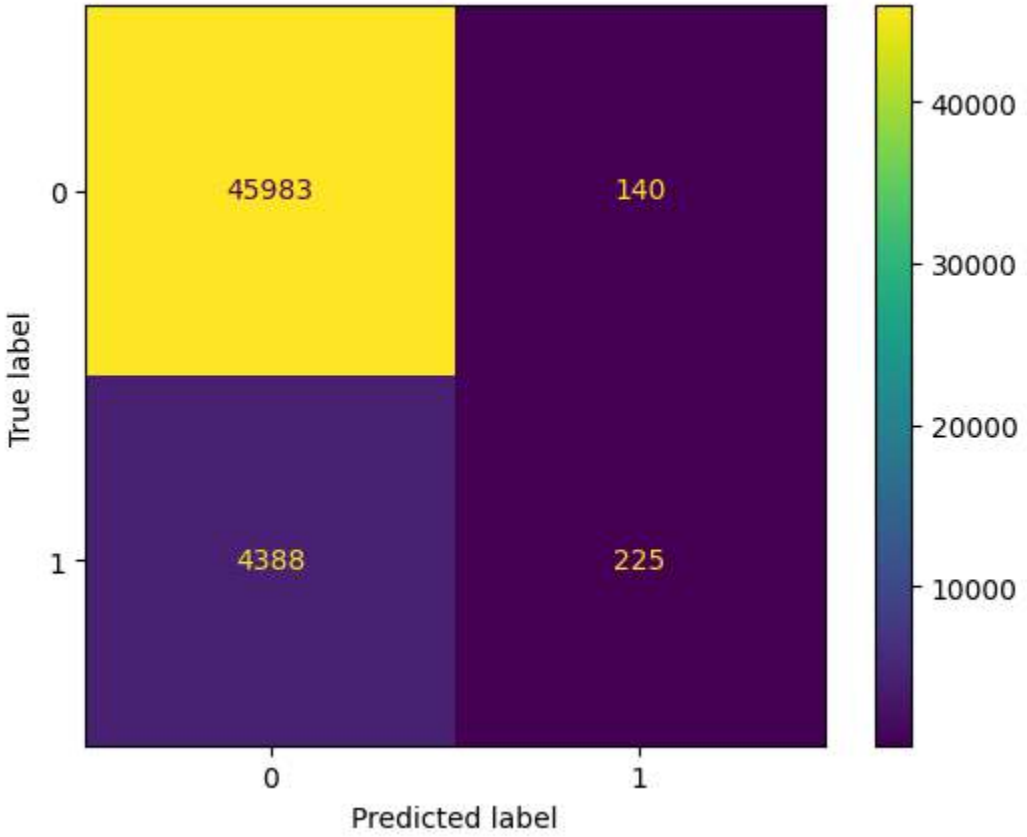
Methods

The base for this project was built using Google Colab, where one can code various AI models in Python. I downloaded the data for the AI to be trained and tested on the Behavioral Risk Factor Surveillance System and then imported it into Google Colab. It contains many different attributes about a person which affect the decision of whether or not someone has heart disease. Some of these attributes include if someone is eating fruits, has high cholesterol, and has high blood pressure. First, I read the dataset in the program with Pandas. After this, the data typically needs to undergo preprocessing so that the AI does not fail while running due to errors or missing values in the data. The reason for this is because 1, the dataset chosen did not have any errors, and 2, a random forest classifier does not need the data to be preprocessed. From here, I sorted the data by separating the first column from the rest, to separate the "answers" from the "evidence" using numpy, which is good for working with arrays. From here, the data set was split into testing and training datasets. The random forest classifier was then created using sklearn and trained on the training dataset. Afterwards, it was tested and the accuracy score was printed out. Then, the random forest classifier's hyperparameters were tuned, using randomly chosen numbers of trees and depths for each tree using Scipy (Virtanen et al.). Sklearn then found the best hyperparameters for the Random Forest Classifier and the hyperparameters and the second accuracy score were printed. Using sklearn again, a confusion matrix was also printed out, and 3 decision trees with a depth of 2 were also printed utilizing graphviz. A bar chart that shows the most influential factors for the random forest classifier was also printed.

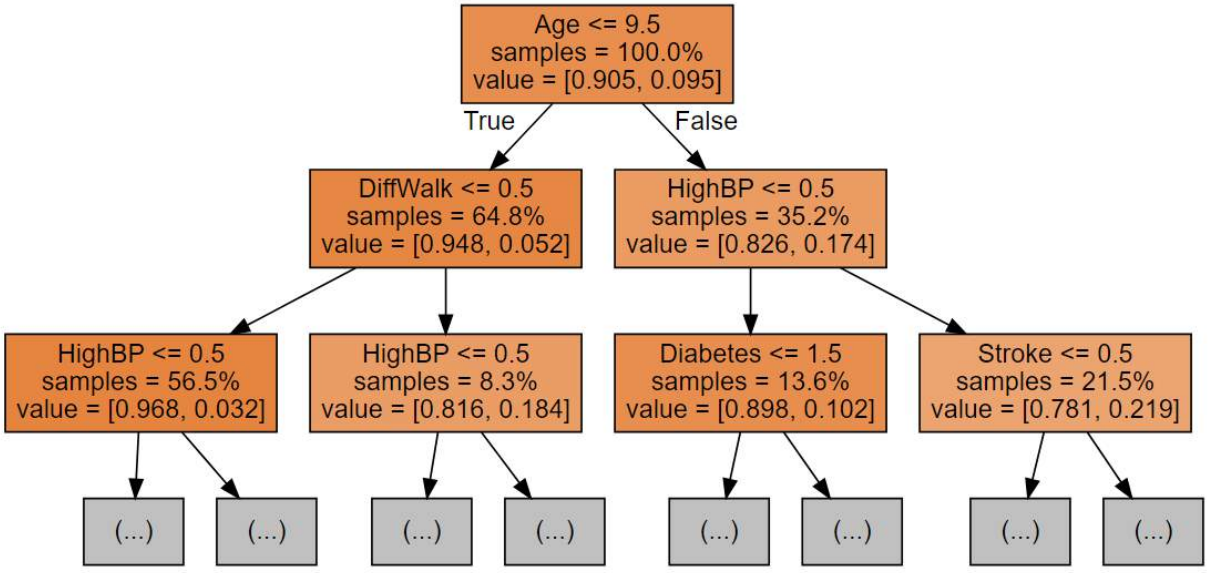
Results

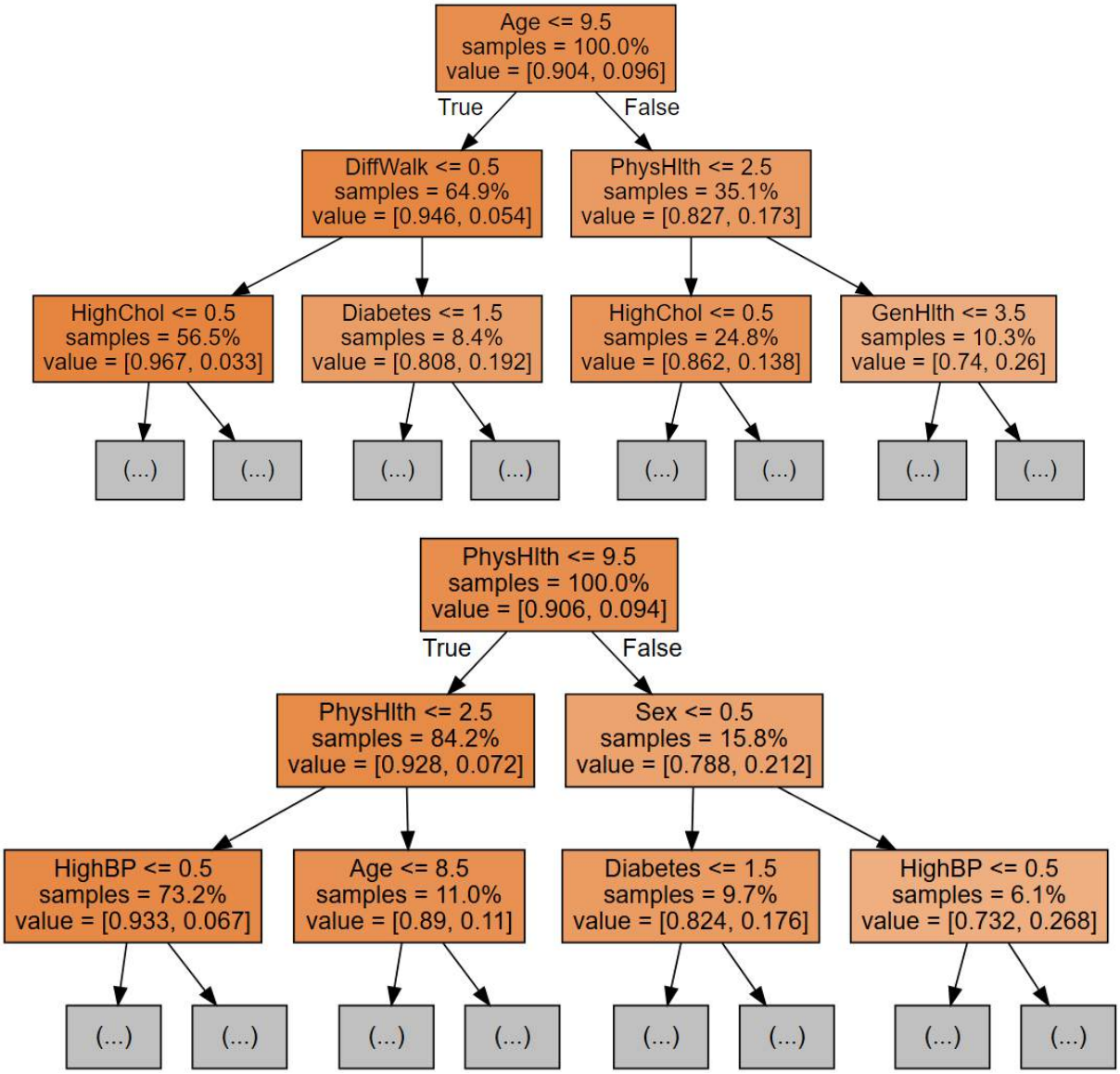
Before the hyperparameters were assigned, the accuracy of the random forest classifier was approximately 90.602%, and after the hyperparameters, the random forest classifier had approximately 91.075%, and the hyperparameters that created this were 469 trees with a depth of

10. The confusion matrix that outputted was this:

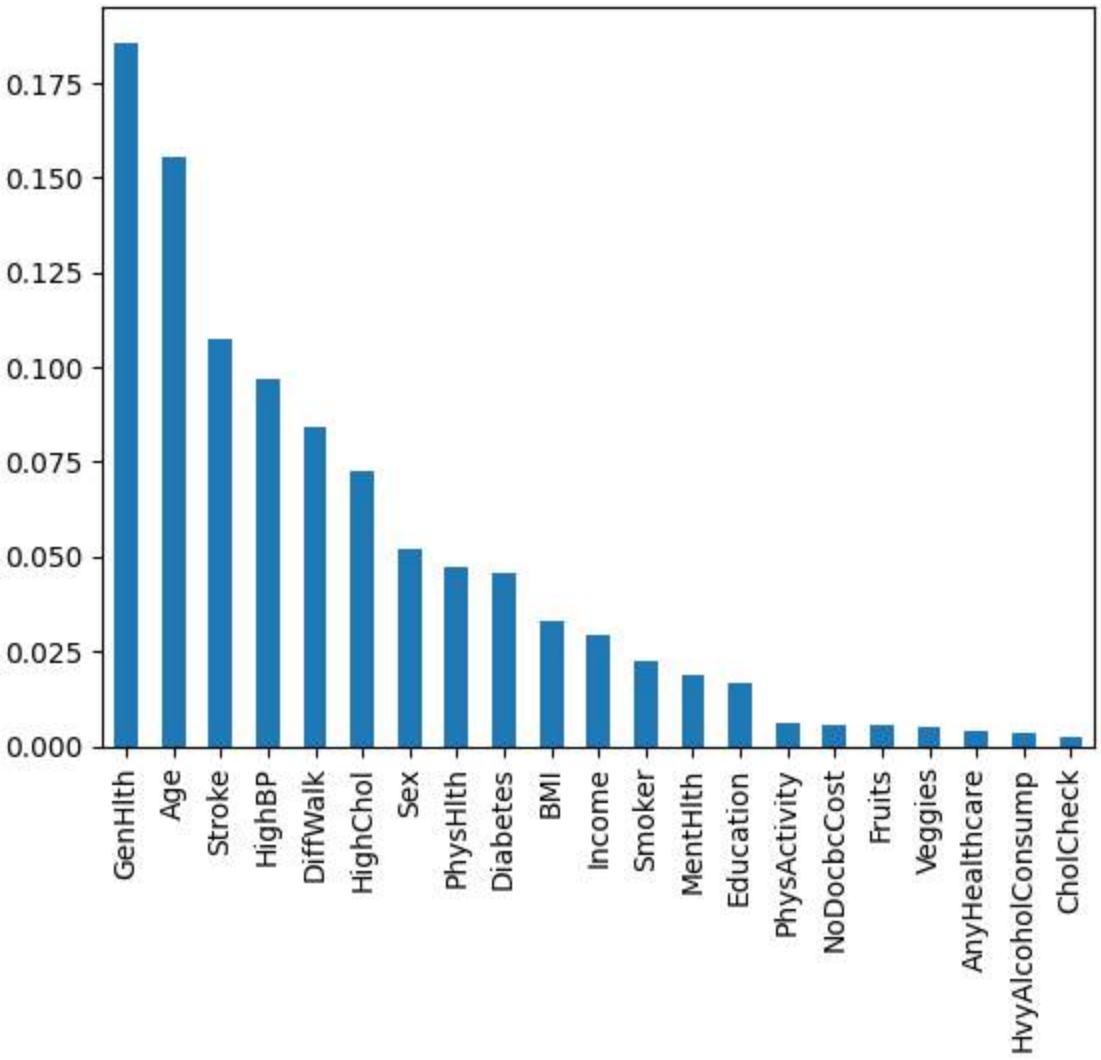


The decision trees that were outputted also looked like this:





The feature importance bar chart looked like this, and shows how much each feature impacted the decision on if someone has heart disease:



Discussion

The accuracy score may show a success for the classifier because it has been shown to have quite a high accuracy at around 90%. However, looking at the confusion matrix, we see that the classifier did not correctly predict many of the cases where someone did have heart disease. This is caused by the massive disparity in the amount of cases with heart disease to cases without heart disease. The random forest classifier was essentially trained and tested on cases that did not have heart disease and of course, is most likely to think that most people did not have heart disease. Yet, the artificial intelligence found that the most important factor that separates someone from someone without heart disease was their general health, their age, and if they have had a stroke. The general health was a self-assessment by the participants in the survey about how they felt they were doing; the age was a number from 1 to 13 where the people would input their age, starting from 18 and each number increase would be 4 years. The stroke column was whether or not the person had a stroke. These top factors are also supported through the decision trees as the decision trees show that age was typically one of the first decisions made in the tree,

indicating its importance to the decision. The specific age that corresponds to heart disease is 56 years and younger. The decision trees reflect that when dividing it between those younger and older than 56, the younger boxes are darker, reflecting more patients 56 and younger with heart disease.

Conclusion

This paper has described an artificial intelligence model using a random forest classifier to determine what factors cause heart disease. The model found that the factors of general health and age were the most important for predicting if someone had heart disease. This information can save lives, as making sure that people know if they have heart disease early can allow them to fight it early as well. This is also a big step in the use of AI in healthcare because AI is a rapidly growing field and if people use it correctly, it can be a major ally in diagnosing patients and finding out if someone has a disease. In the future, people can expand upon this work by creating other artificial intelligence for other diseases like retinal disease or lung disease to try and predict the causes of those.

Works Cited

Dalen, James E., et al. “The Epidemic of the 20th Century: Coronary Heart Disease.” *The American Journal of Medicine*, vol. 127, no. 9, Sept. 2014, pp. 807–12, https://doi.org/10.1016/j.amjmed.2014.04.015.

Heart and Stroke Foundation of Canada. “Types of Heart Disease.” *Heart and Stroke Foundation of Canada*, www.heartandstroke.ca/heart-disease/what-is-heart-disease/types-of-heart-disease.

Rosenson, Robert S., et al. “Cholesterol Efflux and Atheroprotection: Advancing the Concept of Reverse Cholesterol Transport.” *Circulation*, vol. 125, no. 15, Apr. 2012, pp. 1905–19, https://doi.org/10.1161/CIRCULATIONAHA.111.066589.

Schneider, Elizabeth H., et al. “Differential Distribution of Cholesterol Pools across Arteries under High-Cholesterol Diet.” *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*, vol. 1867, no. 12, Dec. 2022, p. 159235, https://doi.org/10.1016/j.bbalip.2022.159235. Accessed 11 Jan. 2023.

Seckeler, Michael D., and Tracey Hoke. “The Worldwide Epidemiology of Acute Rheumatic Fever and Rheumatic Heart Disease.” *Clinical Epidemiology*, vol. 3, Feb. 2011, p. 67, https://doi.org/10.2147/clep.s12977.

Silbernagel, G., et al. “High-Density Lipoprotein Cholesterol, Coronary Artery Disease, and Cardiovascular Mortality.” *European Heart Journal*, vol. 34, no. 46, Sept. 2013, pp. 3563–71, https://doi.org/10.1093/eurheartj/eht343. Accessed 5 Aug. 2022.

Teboul, Alex. “Heart Disease Health Indicators Dataset.” *Www.kaggle.com*, 2022, www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset.

The pandas development team. Pandas-dev/pandas: Pandas. v2.2.3, Zenodo, 20 Sept. 2024, doi:10.5281/zenodo.13819579.

Virtanen, Pauli, et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python.” *Nature Methods*, vol. 17, no. 3, Feb. 2020, pp. 261–72.

World Health Organization. “Cardiovascular Diseases.” *World Health Organisation*, 2024, www.who.int/health-topics/cardiovascular-diseases#tab=tab_1.

To What Extent Does the Ministry of Panchayati Raj Succeed in Implementing Current and Emerging Policies with Regards to the Rural Population of India, Specifically Those Governed By a Gram Panchayat By Parth Vora

Abstract

This paper examines the effectiveness of the Ministry of Panchayati Raj (MoPR) in implementing policies aimed at rural development in India, specifically through Gram Panchayats at the grassroots level. It delves into the decentralized governance structure established by the 73rd and 74th Constitutional Amendment Acts, which introduced a three-tier system of governance and emphasized devolution of power to local self-governments. The analysis focuses on two key schemes: the Rashtriya Gram Swaraj Abhiyan (RGSA) and the Incentivization of Panchayats Scheme, as outlined in the 28th report of the Standing Committee on Rural Development and Panchayati Raj. This paper informs the audience about the RGSA, aimed at strengthening Panchayati Raj Institutions, faced significant issues with the timely release and utilization of funds, highlighting inefficiencies in the MoPR's execution. Inefficiencies in planning and communication across federal, state, and local levels hampered its success, counteracting the potential benefits of decentralized governance. Conversely, the Incentivization of Panchayats Scheme, which includes the National Panchayat Awards, showed more success due to proactive measures and increased funding and awareness as well.
The goals of this paper is to inform the audience about the successes and failures of the MoPR's approach to decentralized governance, provide insights into the challenges of implementing policies, and argue for the importance of efficient fund utilization and responsiveness to policy recommendations. This report analysis highlights how adopting recommendations and addressing systemic inefficiencies can significantly enhance the effectiveness of decentralized rural development policies in India. It concludes by discussing how the MoPR's success in policy implementation is contingent on its responsiveness to recommendations and its ability to ensure efficient fund utilization.

1. An Introduction to Gram Panchayats and the Ministry of Panchayati Raj

1.1 Gram Panchayats and Decentralized Governance in India

Gram Panchayats in India are the typical specimen of Decentralized power within a particular nation. The Decentralization of power is typically defined as the political, administrative or fiscal powers of the central government being distributed amongst smaller, territorially defined authorities or agencies (Richard Crook). Decentralization of power has been broadly categorized into three different subsets; de-concentration, delegation and devolution. The authority and mandate of Gram Panchayats can be viewed from the lens of the latter, while analysing the distribution of power under the 73rd and 74th Amendment Acts of the Indian Constitution.

Devolution is founded on the principle that national, and sub-national agencies share power, wherein the sub-national agencies are granted autonomous legal, financial and/or political authority over an agreed upon territory (Turner and Hulme, 1997). The 73rd and 74th Constitutional Amendment Acts, 1992, detail a tier-based system of Panchayats at the village, intermediate and district levels, and the establishment of Municipalities in urban areas. Under these constitutional amendments, a state is required to devolve appropriate authority and finance upon these bodies, such that they have the ability to develop and implement plans and/or schemes for economic development and social justice. These acts are the fundamental framework for the decentralization of power in India, transferring authority from the central government, down to the states, and eventually local self-government (Dr. Aleya Mousami Sultana).

The 73rd Amendment Act, specifically focuses on the concept of Panchayati Raj, breaking down this system into a series of three tiers. The Gram Panchayat, at the village level, The Intermediate Panchayat at the Taluka Level*, and the Zilla Parishad at the district level. The Gram Panchayat governs a particular village, and is elected into power every five years. It consists of a Sarpanch: the head of the Panchayat, and several (usually up to 9) ward members. A Gram Panchayat is responsible for the cultural, infrastructural and economic development of the village which it governs, while being expected to provide basic services such as but not limited to; education, transportation, healthcare, hygiene etc. The intermediate Panchayat, or the Panchayat Samiti acts as a bridge between several Gram Panchayats, and the district authority, which is the Zilla Parishad. The Zilla Parishad is accountable for the preparation and implementation of developmental plans in the rural areas of a particular district. Its functions are similar to that of a Gram Panchayat, but apply to a wider geographical area.

The Ministry of Panchayati Raj (MoPR) was established in 2004 as a branch of the central government. Its vision is to "attain decentralized and participatory local self-government through Panchayati Raj Institutions." The MoPR functions to advocate for, monitor and implement the 73rd constitutional amendment. It is required to work towards strengthening the administrative infrastructure by providing the appropriate resources and technology to, and facilitating the capacity building measures of said Panchayati Raj Institutions. The entirety of its mandate can be found in part IX of the Indian Constitution, particularly article 243 ZD of Parth IX A pertaining to the District Planning Committees and the Eleventh Schedule which lists 29 subjects to be considered by State Legislatures for devolution to the Panchayats pertaining to the implementation of schemes regarding economic and social development such that said Panchayats function as "Units of Self-Government." However, the effectiveness of the MoPR in carrying out its mandate can come into question. Lack of cooperation, and communication between the central, state and rural, local self-governments can be accredited with dampening the spirit of the 73rd and 74th Amendment Acts, calling into review the true impact of Decentralized governance, and the responsibilities of the MoPR.

1.2. The Standing Committee on Rural Development and Panchayati Raj & the Effectiveness of Decentralized Governance

Decentralized governance, as explained above in the context of Gram Panchayats, is said to be highly effective, in brining authority closer the grassroot problems, thus providing a much greater insight into the needs of the people, eventually leading to good development outcomes (Rajasekhar, Devendra Babu and Manjula, 2018). However, it can be argued that due to a lack of communication, and effective planning from State and federal governments, agencies like the Gram Panchayat are unable to effectively carry out their functions. A lack of united funding, general awareness, and incentive within rural areas about the role of Gram Panchayats essentially counter-acts the promising positives of decentralized governance. Due to a lacklustre approach from the central and state government, especially in aspects such as delayed, and stringent funding, Gram Panchayats grow reluctant to suggest or plan ambitious developmental projects in the first place, further stagnating the role of decentralized government(L.Thirupathi). Considering the Ministry of Panchayati Raj is responsible in ensuring the 73rd Amendment Act is carried out to the right effect, and the Panchayat system is successful in increasing rural development, its efficiency and effectiveness can certainly be called into question.

As of 1983, India employs the system of Standing Committees, that, under their jurisdiction, cover every Ministry/ Department of the Central Government, including the Ministry of Panchayati Raj. A standing committee is expected to, under its mandate: review Bills presented to them by the Speaker/Chairman of the Lok Sabha and/or the Rajya Sabha; review the annual reports on the specific Ministry/Department assigned to them and national policy documents referred to them by the Speaker/Chairman of the Lok Sabha and/or the Rajya Sabha, whereupon reviewing the above, they curate reports entailing their reviews and comments, which are then submitted to the relevant Ministry/Department to be enacted upon. Once these reports have been reviewed and verified by the concerned Ministries/Departments, they are presented to the Lok Sabha/Rajya Sabha. In this manner, the Ministry of Panchayati Raj falls under the jurisdiction of the Standing Committee on Rural Development and Panchayati Raj.

The Standing Committee on Rural Development and Panchayati Raj regularly reviews the functions and policies of the Ministry of Panchayati Raj, creating reports revolving around specific policy-oriented actions, their previous comments, and an analysis of the government's response to said comments. These reports effectively highlight the effectiveness and/or ineffectiveness of the MoPR, and the government's willingness to induce change. This paper will focus primarily on the 28th report(2021-2022) of the Standing Committee: *Action taken on the Observations/Recommendations contained in Twenty Fourth Report on 'Demands for Grants' (2022-23) pertaining to the Ministry of Panchayati Raj.*

1.3 28th Report of the Standing Committee on Rural Development and Panchayati Raj

The 28th report of the Standing Committee reviews the Government's response to the 24th report presented to the Government (Lok Sabha: 17th Lok Sabha and Rajya Sabha), pertaining to the Demands for Grants of the Ministry of Panchayati Raj. Considering that the report analyses

the Government's effectiveness in implementing policy recommendations from a previous report, it can act as adequate evidence in analysing the Government's ability in employing a pro-active response towards rural development.

The 24th report of the Standing Committee is a review of the MoPR's Demand for Grants (2022-23). The Demand for Grants (2022-23) were summarized by the report as follows:

(Rs. in crore)

S.No	Name of the scheme	BE 2022-23
	Plan Scheme	
1.	Rashtriya Gram Swaraj Abhiyan (RGSA)	593.00
2.	Incentivization of Panchayats	50.00
3.	Mission Mode project on E-Panchayats	20.00
4.	Action Research & Publicity	13.00
5.	International Co-operation	0.20
6.	Survey of Villages and Mapping with Improvised Technology in Village Areas (SVAMITVA)	150.00
	Non-Plan	
7.	Secretariat Service	42.37
	Total	**868.57**

Fig1: Demand for Grants (2022-23)

The 24th report is divided into 2 parts, the first part denoting the allocation of funds, and the Ministries justifications of the above allocations, also noting their subsequent plans to adequately use the funds. Part 2, being of greater emphasis in this paper, consists of observations and recommendations made by the Standing Committee with regards to the aforementioned Demands for Grants. On the basis of part 2 of this report, the 28th report dives into the Government's response to these observations/recommendations, after the 24th report was presented to the Lok Sabha and laid in the Rajya Sabha on 16.03.2022.

The 28th report analyses the usage of the allocated funds for the schemes mentioned above, based on the policy recommendations made in part 2 of the 24th report. The report divides itself into four sections, this paper, and the report itself focuses primarily on the following:

Section I: Observations/Recommendations which have been accepted by the government

Section III: Observations/Recommendations in respect of which Replies of the Government have not been accepted by the Committee

This report will draw emphasis primarily on Chapters I and IV of the report. Chapter I denotes the MoPR's initial response to every policy recommendation made in the 24th report, and the Standing Committee's comments and observations of the same. Chapter IV looks specifically at the responses with which the Committee was dissatisfied, and focuses on the lack of an adequate actionable response by the government. This paper will draw greater attention to the points highlighted in Chapter IV of the report, with data observed from Chapter I to be drawn in relation to these points. Drawing data from both these chapters, this report will look to understand the extent to which the MoPR has been deemed effective by the Standing Committee. This report will specifically focus on the following of the aforementioned schemes from the 24th report, and the subsequent policy recommendations (24th Report), responses and comments (28th report) surrounding the same. These focus areas are :

i) Rashtriya Gram Swaraj Abhiyan: The Discrepancy in Approved Plans and Funds Released
ii) Incentivizing Gram Panchayats: Incentivization of Panchayats (National Panchayat Awards) Scheme 2011

The following sections in this paper will each delve into one of the above focus areas, explaining the relevance of said schemes, the policy recommendations made by the Standing Committee with regards to said scheme, and the effectiveness of the MoPR, upon accepting these recommendations in implementing them successfully. In this manner, this paper aims to provide a holistic analysis as to the extent to which the MoPR is successful in implementing current and emerging policies to do with rural development and Gram Panchayats.

2. Rashtriya Gram Swaraj Abhiyan: The Discrepancy in Approved Plans and Funds Released

The Rashtriya Gram Swaraj Abhiyan (RGSA) was launched as an umbrella scheme of the Ministry of Panchayati Raj in 2018, with the fundamental aim of strengthen the powers of the Panchayati Raj System across rural India, in order to achieve the Sustainable Development Goals.** The RGSA extended to all states and union territories of India. It was proposed to be implemented as a core centrally sponsored scheme from the calendar year 2018-19 to 2021-22. The scheme essentially looked to direct more funds towards Gram Panchayats, and other agencies involved in implementing rural development plans, further empowering their capability to influence rural development. The total cost of the scheme was to be Rs. 7255.50 crore, to be distributed amongst the Centre and State governments in a 60:40 ratio for all States except NE and Hilly Territories, for which the distribution ration would be 90:10, and all union territories to be sponsored entirely by the central government.

The funds were to flow in the following way: an initial instalment of 50% of the approved fund from the annual plan, after the deduction of unspent balance from the previous year's budget was to be sent immediately to the Consolidated Fund of the State Government using the Public Financial Management System(PFMS). The next 50% would be released after 60% of the initial funds have been used up. The release of these funds were directly dependent on regular reports to be made by the States on the progress of the scheme's implementation. Upon receiving the Central Government's component of the funds, the State was expected to release said funds to implementing agencies within 15 days. The activities the funds were required to support were:

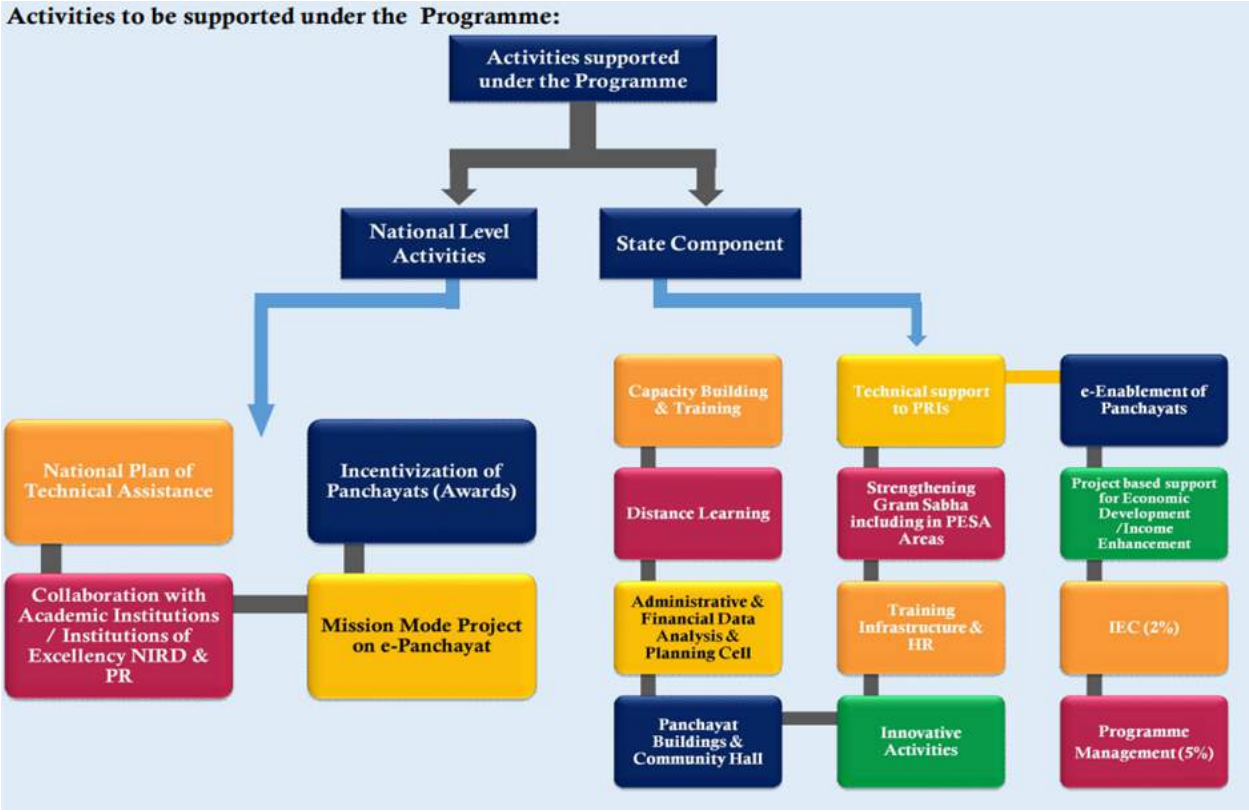


Fig 2: Activities Supported Under the Programme

It is important to note the initial plan as to how the funds were expected to flow for the implementation of the RGSA as the rest of this section will focus on how there was a glaring discrepancy in the plans approved under the scheme, and the funds released, as is mentioned in the Standing Committee's report.

Per the Standing Committee's report, data collected from the ministry itself showed that of 3337.87 crore approved under the plan in 2020-21, only 499,93 crores had been released for use in implementation, and of 4480.22 crore approved in 2021-22, only 518.06 crore had been released. The primary objective of the RGSA was to ensure that the Panchayat government functions efficiently and has the capability to execute developmental plans. However, the State

Governments lacklustre, and untimely approach in releasing funds to Approved Action Plans (AAP), essentially nullified a crucial goal of the RGSA.

The Committee observed the limitations in the Ministry of Panchayati Raj's jurisdiction upon ensuring the timely release of funds with regards to the share of funds contributed by the state government, however, it was of note, that there was a clear, untimely release of funds through the years 2019-2022, to Union Territories which fall under the RGSA, such as the Andaman and Nicobar Islands, where 100% of the funds are to be contributed by the central government, falling directly under the mandate of the Ministry of Panchayati Raj.

On April 2nd 2018, PRI members of the Sundargarh Panchayat, Baratang Island, ANI, declared a hunger strike, protesting the lack of released funding to their approved infrastructural development plans, specifically, the funds required to build a bridge in Karah Nallah, and three new village roads. Being a Union territory, funding to the Panchayat is provided entirely by the central government, and this incident clearly depicted not only the lack of, and stagnation of rural development due to a discrepancy in plans approved and funds released, for which the MoPR is directly responsible, but a discontent within the Panchayats themselves, in being unable to execute any form of rural development due to a distinct lack of funding, and highly delayed transfer of funds from the centre down to the grassroots.

The ministry in their reply to the initial comments made by the Standing Committee detailed that funds released on their behalf were limited due to the unspent balance available to spend by the State or Union territory; referring specifically to the Andaman and Nicobar Islands, with respect to significant unspent balance from the previous year being available for infrastructure spending, and thus the Ministry itself was not to release any further funds, considering the fact that funds were only to be contributed after deducting the un spent balance from the previous year's contribution. Thus, in the Ministry's response to the comments made by the Standing Committee, they attributed the discrepancy in approved and released funding to unspent funds being available to use, in both the case of State contributions and contributions to have been made by the Centre with regards to the RGSA.

However, this then brings to light the severe underutilization of available funding by Panchayats at the grassroot level, with the Standing Committee calling into question the Ministry's role as a facilitator in implementing these funds and thus enabling the adequate use of these funds, in which they have clearly been unsuccessful. The Ministry was urged to undertake a more solemn approach in ensuring the release of and apt utilization of funds at the grassroot, specifically under the Rashtriya Gram Swaraj Abhiyan, considering the RGSA was set to be pivotal in several facets of rural development, and as of that moment, the Ministry had had very limited success in fulfilling their role in facilitating rural development.

Next, we are to look at the Ministry's role and success in incentivising Gram Panchayats, and the Standing Committees take on and analysis of the same.

3. Incentivization of Panchayats: Allocation of Funds and Number of Awards

The Ministry of Panchayati Raj, under the Incentivisation of Panchayats Scheme, recognized as a key component of the revamped Rashtriya Gram Swaraj Abhiyan, introduced the National Panchayat Awards as a means to felicitate and thus incentivise the best performing Panchayats across the nation. These awards, although established in 2011 and delivered annually since then, were revamped and re-introduced in 2022 to be aligned with the 9 Localizations of Sustainable Development Goals outlined by the Ministry of Panchayati Raj in their Expert Report on the Localization of Sustainable Development Goals in PRIs. They were intended to act as a means for Panchayats to be incentivised to achieve these LSDGs by assessing their progress in achieving them in lines with the 2030 agenda, on a competitive basis, such that friendly competition in awarding the best performing Panchayats would incentivise efficiency in rural development.

The awards now follow a pyramidal structure spanning across the Panchayat, District and State/UT level. This pyramid structure can be simplified to the following, wherein the awards are distributed across several stages of self-local government:

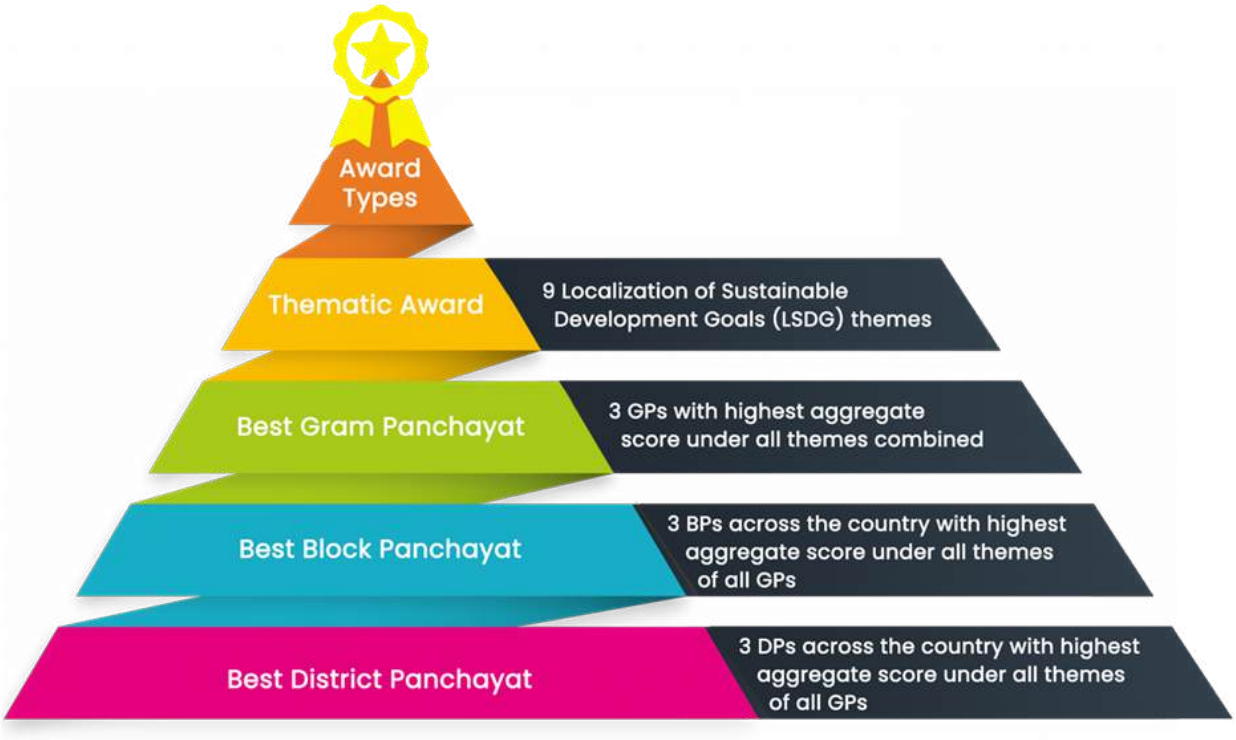


Fig 3: Award Pyramid

The awards at each of these levels would then be distributed under two categories:

a) Deen Dayal Upadhyay Panchayat Satat Vikar Puraskar (Based on performance with regards to each individual theme)

b) Nanaji Deshmukh Sarvottam Panchayat Satat Vikas Puraskar (Aggregate performance under all of the themes)

Along with each award, Gram Panchayats win cash prizes ranging from 50-75 lakhs for being recognized in any of the aforementioned categories.

Having established the National Panchayat awards as a key functionary of the Incentivisation of Panchayats Scheme undertaken by the MoPR, the standing committee went to detail the following in terms of its implementation in their report:

The standing committee believed that the funding dedicated to the "Incentivization of Panchayats" had grown increasingly stagnant hindering its possible further effectivity. Since the financial year 2020-21; i.e., the last three years, the budget allocated to the incentivization of panchayats has moved from Rs. 47 crores to Rs, 50 crores, an overall increase of only Rs. 3 crores. This is of note considering the expenditure under this scheme has almost always elapsed the budget allocated. This speaks to the fact that a progressive increase in funding would have helped contribute to a more expansive and effective implementation of this scheme. The standing committee believes, the Ministry should have noted the need to increase funding towards this scheme to magnify its positive effects, rather than let it stagnant over a period of three years. As detailed above, a number of awards are given out in recognition and appreciation of Gram Panchayats and their efforts to improve public services and rural development. However, even after seeing the success of this scheme in incentivizing gram panchayats, the Ministry chose not to increase the number of awards given out to the said Panchayats. The Standing Committee believed that increasing the number of awards given out would widen the demographic for effective incentivisation, and therefore the original scheme should have been revised to include a greater number of awards upon seeing the initial success of the scheme.

Further the Standing Committee believed there to be a lack of awareness about the awards, hindering possible further incentivisation across a larger scale and a greater number of Panchayats. The standing committee expected the Ministry to engage in greater efforts to prevent the impact and effectiveness of an initially successful scheme from growing stagnant in its results due to a reluctance to increase its scale and funding.

In response to these initial observations made by the Standing Committee the Ministry of Panchayati Raj stood firmly in the belief that they had indeed employed extensive efforts in raising awareness about the awards, via print, radio and other forms of multimedia and had collaborated successfully with both States and Union Territories in ensuring the same. They emphasized on the fact that the awards are given out annually on the 24th of April, nationally recognized as National Panchayati Raj Day (NPRD): a widely publicized national level event, denoted by majority Gram Panchayats across the nation. The NPRD function is a considerably large event attended by a considerable number of Panchayats across the nation, thus garnering distinct publicity for the awards.

In terms of the funding allocated to the Incentivisation Scheme, the Ministry specified that the current allocation of Rs. 50 crores was decided under the revamped Rashtriya Gram

Swaraj Abhiyan, which the Ministry believed was one of its central components, and added to this notion by stating that several initiatives were in the process of being undertaken by a series of other Ministries and Departments to sponsor these awards, allowing them to direct more funds towards them, and upscale them, recognising a greater number of Gram Panchayats, increasing the impact of the incentivization scheme.

This response from the Ministry saw a positive reply by the Standing Committee in the report, noting the Ministry's efforts in obtaining more sponsors to further fund the National Panchayat Awards, aligning with the initial recommendations made by the committee. In contrast to the aforementioned discrepancy in funding with regards to the RGSA, we see the success the MoPR can have at positively influencing rural development if they take an active stance on improving existing policies with regards to Gram Panchayats. Both, the success of the Incentivisation Scheme and the positive response from the Standing Committee in their report stand testament to the fact that the MoPR, when efficient in its implementation of policy, can improve the efficiency and effectiveness of Gram Panchayats.

4. Conclusion

The Ministry of Panchayati Raj has seen both distinct failures and certain successes in adequately implementing policies to do with facilitating rural development via Gram Panchayats, as demonstrated by the 28th Report of the Standing Committee on Rural Development and Panchayati Raj. This paper has noted the comments and policy recommendations made by the Standing Committee with regards to specific policy measures undertaken by the Ministry and has then gone on to see the government's response to the same, thus developing a wholistic view on the extent to which the Ministry is able to effectively carry out policy measures they undertake. With regards to the first Scheme analysed by this paper, that being the Rashtriya Gram Swaraj Abhiyan, it was clear that although the Ministry undertook an ambitious project in aiming to hasten development at the grassroots, the inefficiency in delivering approved funding to Panchayats in order to execute this development made it increasingly difficult for any visible outcome of this scheme to materialise in terms of rural development. The comments made by the Standing Committee with regards to the same, were met by vague, and unaccepting replied from the government, displaying a lacklustre approach in trying to improve existing policy measures to increase their effectiveness. This gaping inadequacy in the Ministry's implementation of the RGSA showed the limited impact its policy measures can truly have when carried out inefficiently.

In contrast, the second Scheme analysed by this report, that being the Incentivisation of Panchayats Scheme, presented opposing observations. A positive response from the government in actively trying to increase funding towards and awareness about the National Panchayat Awards saw a further positive notation by the Standing Committee. The approach by the Ministry under this Scheme, to actively improve and develop the existing policy and take an active stance in responding to the comments made by the Standing Committee saw an increase in the successful implementation of the policy. This highlighted how, if and when the Ministry

adopts new policy measures, and takes into account policy recommendations made by the Standing Committee they are able to garner increased success in improving the efficiency of Gram Panchayats, thus contributing to enhanced rural development.

We can conclude that although there are several policy measures undertaken by the Ministry of Panchayati Raj, the extent of their success can be related to the approach taken by the Ministry to ensure their efficiency. If the Ministry is more open to adopting policy recommendations made by organizations like the Standing Committee, policies implemented by the Ministry are likely to see greater success in impacting Gram Panchayats and Rural Development.

Works Cited

Standing Committee on Rural Development and Panchayati Raj. 28th Report Standing Committee on Rural Development and Panchayati Raj. 8 Mar. 2022. Lok Sabha, Retrieved 30 Oct. 2023, https://loksabhadocs.nic.in/lsscommittee/Rural%20Development%20and%20Panchayati%20Raj/17_Rural_Development_and_Panchayati_Raj_28.pdf.

Ministry of Panchayati Raj. National Panchayat Awards. n.d., https://panchayataward.gov.in/.

S. "Lack of Funds and Slow Developmental Works: Sundargarh Panchayat PRI Members and Public Begin Indefinite Hunger Strike." Andaman Sheekha, 2 Apr. 2018, https://www.andamansheekha.com/61317/.

Government of India. Rashtriya Gram Swaraj Abhiyan. n.d., https://rgsa.gov.in/index.htm.

Crook, Richard. Decentralization and Good Governance. n.d., Forum of Federations, https://www.forumfed.org/libdocs/IntConfFed02/StG-Crook.pdf.

Turner, Mark, and David Hulme. Governance, Administration and Development. Palgrave Macmillan, 1997.

Sultana. Local Self-Government Course-424. n.d.

"30th Anniversary of the 73rd and 74th Amendments." Drishti IAS, n.d., https://www.drishtiias.com/daily-updates/daily-news-analysis/30th-anniversary-of-the-73rd-and-74th-amendments.

Misra, N. "3-Tier Structure of Panchayati Raj." Your Article Library, 27 Oct. 2015, https://www.yourarticlelibrary.com/panchayati-raj-2/3-tier-structure-of-panchayati-raj/66692.

Ministry of Panchayati Raj. Organization Structure. n.d., https://panchayat.gov.in/organization-structure/.

Standing Committee on Rural Development and Panchayati Raj. 24th Report of the Standing Committee on Rural Development and Panchayati Raj. 16 Mar. 2022.

Thirupathi, L. "Democratic Decentralization and Devolution of Powers at the Grassroots Level Democracy: Issues, Challenges, and Implications." International Journal of Advanced Research, vol. 9, May 2021, pp. 947-951, ISSN 2320-5407, www.journalijar.com.

Comparative Review on Traditional Treatments and Genetic-Based Therapies in Treating Pancreatic Cancer By Jessica Li

Abstract

Pancreatic cancer is an extremely dangerous condition due to its ability to quickly spread throughout the body, therefore, it is crucial to develop personalized treatments that can both detect the cancer early and prevent it from spreading throughout the body. Traditional treatments like surgery, chemotherapy, and radiation offer limited success, often extending life only slightly Whilet, emerging genetic therapies like CRISPR-Cas9, CAR T-cell therapy, and cancer vaccines target specific mutations present in most pancreatic cancer cases. These treatments aim to reduce side effects and improve survival by focusing on cancer cells while sparing healthy tissue. However, challenges like ensuring safety, and efficacy. I will write a comprehensive literature review summarizing the current research in pancreatic cancer and its future perspectives on treatments. Overall, this project's main goal is to spread awareness of pancreatic cancer by discovering the different types of treatments through research.

Introduction

As of 2023, the five-year survival rate of pancreatic cancer in the United States is only about 12% (Post). Pancreatic cancer is an extremely fatal disease with a poor prognosis and limited treatments. As people often call it the “silent killer,” the symptoms of pancreatic cancer typically remain hidden until the tumor has advanced within the body, making early detection challenging and limiting the effectiveness of treatments (Fairley et al.). However, when pancreatic cancer can finally be detected through methods such as imaging scans, it is often too late, as the cancer would’ve already rapidly spread within the body with the assistance of neurons and surrounding cells through processes such as perineural invasion (Chen et al.). Due to the severity of this condition, this paper will further explore pancreatic cancer treatments, as traditional ones are generally ineffective and thus require us to continue to look for alternative treatment options.

Background of Pancreatic Cancer

Located behind the stomach, the pancreas plays a crucial role in digestion and blood sugar regulation. As it is covered by other organs and located deep within the body, the most common type of pancreatic cancer, exocrine pancreatic cancer, often develops silently, masking all the symptoms only until the cancer has already spread to surrounding vital organs (Johns Hopkins). As symptoms are usually not prominent enough in the early stages as they can often be mistaken for other illnesses, it is crucial to be able to identify possible symptoms caused by pancreatic cancer (Krech and Walsh). Some of the most common symptoms may include weight loss, jaundice, and issues with blood sugar levels, and considering the pancreas’ location within the body may include abdominal pain (Krech and Walsh).

Pancreatic cancer can severely affect the quality of life of patients through physical symptoms as well as emotional and psychological effects of the cancer itself. The symptoms, such as abdominal pain, weight loss, fatigue, and jaundice can severely limit daily activities and reduce the patient's quality of life ough physical symptoms as well as emotional and psychological effects of the cancer itself (Cipora et al.). In addition, a malfunctioning pancreas can lead to digestive issues such as hindering the stomach's ability to break down food due to the lack of enzymes produced by the pancreas (*Pancreatic Cancer Treatment - NCI*). This leads to malnutrition, which can be a factor that contributes to weight loss. Furthermore, the prognosis of pancreatic cancer is often poor, which can lead to significant emotional and psychological distress, including anxiety, depression, and a sense of hopelessness (Cipora et al.). The aggressive nature of the disease may lead to a rapid decline in health, causing strain on both the patient and their family members.

There are three main types of pancreatic cancers, exocrine, neuroendocrine, and benign precancerous lesions, each with its own subcategories (Johns Hopkins). The most common type of pancreatic cancer is exocrine pancreatic cancer. There are many variations of exocrine pancreatic cancer and in total, they make up more than 95 percent of pancreatic cancers, as the majority of the pancreas is composed of exocrine cells (Johns Hopkins). This subtype of pancreatic cancer is in the exocrine cells, which are cells that make up the exocrine gland and ducts of the pancreas and produce enzymes important to digestion. The exocrine glands help to create enzymes to break down carbohydrates, fats, proteins, and acids (*Pancreatic Neuroendocrine Tumors (Islet Cell Tumors) Treatment - NCI*). The second type of pancreatic cancer is called neuroendocrine pancreatic cancer (Johns Hopkins). This type of cancer develops from cells in the endocrine gland of the pancreas. The endocrine gland produces hormones, insulin, and glucagon into the bloodstream to help regulate the level of blood sugar. Endocrine pancreatic cancer is extremely rare and is only less than 5 percent of all pancreatic cancer cases. The last major category of pancreatic cancer is benign precancerous lesions (Johns Hopkins). The benign precancerous lesions are cysts, abnormal noncancerous growth filled with liquid or a semisolid substance, that grows in the pancreas. Some can be precursors to pancreatic cancer including intraductal papillary mucinous neoplasm or IPMNs, which are cystic neoplasms of the pancreas that grow within the pancreatic ducts and produce mucin (Puckett et al.). Tumors are often found when a patient is being scanned for unrelated medical reasons. This may be surgically removed depending on the location and condition to see if it will become malignant.

In addition, due to its ability to remain hidden, it is also important to be able to recognize some of the risk factors that will increase the likelihood of developing pancreatic cancer. Pancreatic cancer risk factors include genetic, lifestyle, and medical factors. Mutations in genes, such as BRCA1, BRCA2, and KRAS, as well as family history significantly increase the risk. BRCA2 mutations are found in about 5–10% of cases, and BRCA1 mutations are found in about 1% (Petrucelli et al.). However, in certain populations that are geographically or culturally distinct, the presence of BRCA mutations is higher, a phenomenon called founder mutations (Petrucelli et al.). Founder mutations are genetic alterations observed in a population that

originated from a small group of ancestors and are often passed down through generations due to limited genetic diversity. KRAS gene mutations play a significant role in the development and progression of pancreatic cancer, particularly in pancreatic ductal adenocarcinoma, which accounts for the majority of pancreatic cancer cases (Huang et al.). KRAS is a key gene involved in regulating cell growth, and mutations in this gene lead to uncontrolled cell division, contributing to tumor formation. In fact, over 90% of pancreatic cancers involve a mutation in the KRAS gene, making it one of the most commonly mutated genes in this type of cancer (Huang et al.). Lifestyle factors like smoking, heavy alcohol consumption, and obesity are the main causes of pancreatic cancer as well (*Pancreatic Cancer Treatment - NCI*). Medical conditions like chronic pancreatitis, long-standing diabetes, and metabolic syndrome also raise the risk of pancreatic cancer (*Pancreatic Cancer Treatment - NCI*). Due to the genetic risk factors of pancreatic cancer, it has brought attention to the importance of the treatments, more specifically, the genetic treatments.

Traditional treatment and genetic treatments

Pancreatic cancer treatment options vary based on the stage and specific characteristics of the cancer. Stage 0 of the cancer typically involves abnormal cells that have a possibility of becoming cancerous. Stages I and II indicate localized cancer, which is when the tumor has grown outside the pancreas area. During these stages, the cancer can be treated with surgery and possibly chemotherapy or radiation. Stages III and IV involve more extensive spread, with Stage III being locally advanced cancer, where the tumor has spread to nearby major blood vessels or nerves and possibly nearby lymph nodes but not to distant sites. Finally, Stage IV is metastatic cancer, where the cancer has spread to distant organs such as the liver, lungs, or peritoneum (*Pancreatic Cancer Stages*)**.**

Traditional treatments include surgery, chemotherapy, and radiation therapy. More personalized treatments include vaccines and medicines. Some of the common side effects of chemotherapy may include hair loss, loss of appetite, nausea and vomiting, and constipation (Gutteridge). Compared to the minimized side effects, genetic treatments may be a better option for cancer treatments, especially for fast-spreading cancers such as pancreatic cancer. Genetic treatments are much more tailored and therefore can be much more effective and have fewer side effects than those one-size-fits-all treatments. Traditional treatments like chemotherapy are mainly used to prevent the division of cancerous cells by killing them, which can lead to side effects such as nausea, hair loss, and fatigue due to their impact on healthy cells (Gutteridge). In contrast, genetic therapies target specific cancer mutations, often resulting in fewer side effects like problems in the skin and liver as they focus primarily on cancer cells.

Typically, traditional treatments such as surgery, chemotherapy, and radiotherapy, have lower success rates due to the aggressive nature of the cancer and its tendency to remain hidden during the early stages. With the traditional treatments, the five-year survival rate for all stages of pancreatic cancer is around 5-10%, only with certain cases having slightly better outcomes at about 15-20% (Yue et al.). Surgery for pancreatic cancer often has limited effectiveness, as it's

challenging to remove all cancerous tissue due to the pancreas's deep location, and it frequently leads to non-specific side effects like digestive issues or infection risks (Yue et al.). Additionally, survival rates remain low, even post-surgery, because pancreatic cancer is typically detected at a late stage when it has already spread, reducing the long-term benefits of surgical intervention. Similarly, chemotherapy and radiation may offer limited survival benefits, often extending life by a few months rather than achieving long-term remission (Yue et al.). Some of the outcomes for patients who receive traditional treatments include limited efficacy, non-specific side effects, and shortened survival rates. On the other hand, personalized genetic treatments have proven to have a much more positive outcome than traditional therapies. For example, some positive outcomes include targeted efficacy, reduced side effects, and improved outcomes and survival.

CRISPR-Cas9 is a cutting-edge gene-editing technology that holds promise for treating pancreatic cancer by targeting and correcting specific genetic mutations that drive the disease (Redman et al.). This technique involves using a guide RNA to direct the Cas9 enzyme to a specific location in the DNA, where it makes a precise cut. In pancreatic cancer, CRISPR-Cas9 can potentially be used to edit or disable genes like KRAS, which is commonly mutated in these tumors (Redman et al.). By targeting these mutations, CRISPR-Cas9 could help stop the growth and spread of cancer cells, offering a more personalized and effective treatment. However, this approach is still in the experimental stages, and more research is needed to ensure its safety and efficacy in humans.

CAR T-cell therapy is an innovative immunotherapy that involves modifying a patient's T cells to recognize and attack cancer cells more effectively (Sterner and Sterner). In the context of pancreatic cancer, this treatment works by extracting the patient's T cells, genetically engineering them to express chimeric antigen receptors (CARs) that specifically target antigens on the surface of pancreatic cancer cells, and then reinfusing these modified T cells back into the patient (Sterner and Sterner). Once inside the body, the CAR T-cells seek out and destroy cancer cells. Although CAR T-cell therapy has shown success in treating certain blood cancers, its application in solid tumors like pancreatic cancer is still under investigation. In this study, CAR T-cell therapy led to long-term remission in 40% of patients with DLBCL, a type of aggressive lymphoma (Sterner and Sterner). Challenges such as the dense tumor microenvironment and potential toxicity need to be addressed to improve its effectiveness against pancreatic cancer.

Vaccines for pancreatic cancer aim to stimulate the immune system to recognize and fight cancer cells. Unlike traditional vaccines that prevent diseases, these therapeutic vaccines are designed to treat existing cancer by inducing an immune response against tumor-specific antigens (Melief et al.). One approach involves using vaccines that target proteins commonly expressed in pancreatic cancer cells, such as mutant KRAS. Another strategy includes using dendritic cell vaccines, where a patient’s cells are exposed to cancer antigens in the lab and then reintroduced into the body to trigger a robust immune response (Melief et al.) Although still in clinical trials, cancer vaccines offer a promising avenue for complementing other treatments, potentially improving survival rates and reducing the likelihood of cancer recurrence in patients with pancreatic cancer.

Conclusion

Pancreatic cancer, particularly the exocrine type, is a highly aggressive and often fatal disease with a poor prognosis. Traditional treatments, including surgery, chemotherapy, and radiation therapy, generally have low success rates due to the cancer's tendency to be diagnosed in advanced stages. Newer, personalized treatments like targeted therapies and cancer vaccines offer more precise treatment options, potentially improving outcomes and reducing side effects compared to traditional approaches. The use of genetic advancements in cancer treatment is significant because it allows for more personalized, precise therapies that target the specific mutations driving a person’s cancer. By focusing on these unique genetic markers, treatments like targeted therapy and immunotherapy can attack cancer cells more effectively while minimizing damage to healthy cells, which can reduce side effects and improve patient outcomes. This approach not only enhances treatment efficacy but also offers a pathway to develop therapies for cancers that were previously difficult to treat with conventional methods.

Works Cited

Chen, Shu-Hai, et al. “Perineural Invasion of Cancer: A Complex Crosstalk between Cells and Molecules in the Perineural Niche.” American Journal of Cancer Research, vol. 9, no. 1, Jan. 2019, p. 1.

Cipora, Elżbieta, et al. “Quality of Life in Patients with Pancreatic Cancer—A Literature Review.” International Journal of Environmental Research and Public Health, vol. 20, no. 6, Mar. 2023, p. 4895. pmc.ncbi.nlm.nih.gov, https://doi.org/10.3390/ijerph20064895.

Fairley, Kimberly J., et al. “‘Invisible’ Pancreatic Masses Identified by EUS by the ‘Ductal Cutoff Sign.’” Endoscopic Ultrasound, vol. 8, no. 2, 2019, pp. 125–28. PubMed Central, https://doi.org/10.4103/eus.eus_49_15.

Gutteridge, Winston E. “EXISTING CHEMOTHERAPY AND ITS LIMITATIONS.” British Medical Bulletin, vol. 41, no. 2, Jan. 1985, pp. 162–68. Silverchair, https://doi.org/10.1093/oxfordjournals.bmb.a072044.

Huang, Lamei, et al. “KRAS Mutation: From Undruggable to Druggable in Cancer.” Signal Transduction and Targeted Therapy, vol. 6, no. 1, Nov. 2021, pp. 1–20. www.nature.com, https://doi.org/10.1038/s41392-021-00780-4.

Johns Hopkins. Pancreatic Cancer Types. 28 May 2024, https://www.hopkinsmedicine.org/health/conditions-and-diseases/pancreatic-cancer/pancreatic-cancer-types.

Krech, Ruth L., and Declan Walsh. “Symptoms of Pancreatic Cancer.” Journal of Pain and Symptom Management, vol. 6, no. 6, Aug. 1991, pp. 360–67. ScienceDirect, https://doi.org/10.1016/0885-3924(91)90027-2.

Melief, Cornelis J. M., et al. “Therapeutic Cancer Vaccines.” The Journal of Clinical Investigation, vol. 125, no. 9, Sept. 2015, pp. 3401–12. www.jci.org, https://doi.org/10.1172/JCI80009.

Pancreatic Cancer Stages | Staging Pancreatic Cancer. https://www.cancer.org/cancer/types/pancreatic-cancer/detection-diagnosis-staging/staging.html. Accessed 22 Nov. 2024.

Pancreatic Cancer Treatment - NCI. 27 Sept. 2024, https://www.cancer.gov/types/pancreatic/patient/pancreatic-treatment-pdq. nciglobal,ncienterprise.

Pancreatic Neuroendocrine Tumors (Islet Cell Tumors) Treatment - NCI. 1 Nov. 2024, https://www.cancer.gov/types/pancreatic/patient/pnet-treatment-pdq. nciglobal,ncienterprise.

Petrucelli, Nancie, et al. “BRCA1- and BRCA2-Associated Hereditary Breast and Ovarian Cancer.” GeneReviews®, edited by Margaret P. Adam et al., University of Washington, Seattle, 1993. PubMed, http://www.ncbi.nlm.nih.gov/books/NBK1247/.

Post, Erin. “Five-Year Pancreatic Cancer Survival Rate Increases to 12%.” Pancreatic Cancer Action Network, 12 Jan. 2023, https://pancan.org/news/five-year-pancreatic-cancer-survival-rate-increases-to-12/.

Puckett, Yana, et al. “Intraductal Papillary Mucinous Neoplasm of the Pancreas.” StatPearls, StatPearls Publishing, 2024. PubMed, http://www.ncbi.nlm.nih.gov/books/NBK507779/.

Redman, Melody, et al. “What Is CRISPR/Cas9?” Archives of Disease in Childhood - Education and Practice, vol. 101, no. 4, Aug. 2016, pp. 213–15. ep.bmj.com, https://doi.org/10.1136/archdischild-2016-310459.

Sterner, Robert C., and Rosalie M. Sterner. “CAR-T Cell Therapy: Current Limitations and Potential Strategies.” Blood Cancer Journal, vol. 11, no. 4, Apr. 2021, pp. 1–11. www.nature.com, https://doi.org/10.1038/s41408-021-00459-7.

Yue, Qingxi, et al. “Natural Products as Adjunctive Treatment for Pancreatic Cancer: Recent Trends and Advancements.” BioMed Research International, vol. 2017, Jan. 2017, p. 8412508. pmc.ncbi.nlm.nih.gov, https://doi.org/10.1155/2017/8412508.

The New Deal: A Raw Deal for African Americans By Julian Zhang

Introduction

As the Roaring Twenties ended, the U.S. endured the most severe economic recession in the country's history, the Great Depression. With poverty and unemployment rates rising, American citizens looked to the government for a solution. Franklin Delano Roosevelt introduced the idea of the New Deal in 1932 during his presidential campaign against Herbert Hoover. After Roosevelt won the presidency, Congress passed the New Deal in 1933. The New Deal included many projects and policy changes designed to relieve the suffering of the American populace, reform the financial system, and recover the economy. However, the New Deal also exacerbated racial inequalities through segregated housing, discriminatory job programs, and the exclusion of minorities from Social Security.

Policies and Practices

The New Deal's housing policies worsened racial segregation and inequality by policy and practice. Policies enacted by the Federal Housing Administration (FHA), established in 1934, often denied loans to Black Americans. John Brueggemann, a professor of sociology at Skidmore College, writes, "The FHA...distributed low interest loans, but often turned down blacks as potential recipients."[2] Excluding African Americans from loans effectively reinforced segregation as it confined African Americans to overcrowded and under-resourced neighborhoods, while white families were able to move into areas with better housing and schools. Additionally, the practice of leveling Black neighborhoods to erect housing projects damaged Black communities. Mary-Elizabeth Murphy, a history professor at Eastern Michigan University, notes, "When the PWA constructed black housing projects, they engaged in slum clearance by razing Black neighborhoods."[3] Paul Moreno, a professor of history at Hillsdale College, addresses the damage this caused: "Urban renewal in New York showed a similar tendency to overrun black neighborhoods and to promote economic investments that did not benefit blacks."[4] Targeting Black neighborhoods for development projects led to many families being displaced against their will. These destructive actions, as well as denying mortgages for African Americans worsened housing inequality and thus deepened racial divides in America.

While the New Deal aimed to help Americans to recover from the Great Depression, it heightened racial inequalities through systemic racial biases in New Deal policies such as the Social Security Act (SSA) of 1935. Enacted by President Roosevelt, the SSA provided financial assistance to certain demographics, such as the unemployed and the elderly, but excluded

---

[2] John Brueggemann, "Racial Considerations and Social Policy in the 1930s," Social Science History 26, no. 1 (2002): 149, JSTOR.

[3] Mary-Elizabeth B. Murphy, "African Americans in the Great Depression and New Deal," in *Oxford Research Encyclopedia of American History*, 9, last modified November 19, 2020, https://oxfordre.com/americanhistory/view/10.1093/acrefore/9780199329175.001.0001/acrefore-9780199329175-e632.

[4] Paul Moreno, "An Ambivalent Legacy: Black Americans and the Political Economy of the New Deal," *The Independent Review* 6, no. 4 (2002): 528, JSTOR.

farmers and domestic workers, common occupations for African Americans. Murphy notes, "When FDR signed the Social Security Act into law in 1935, it deemed farmers and domestics ineligible, which meant that 87 percent of all Black women and 55 percent of all African American workers were excluded."[5] Moreover, granting states control over these programs led to further discrimination, especially in the South. As Jesse Thomas, a successful African American educator in the early 1900s, wrote in the African-American academic journal *Opportunity*, "Many white people in the South are dogmatically opposed to Negroes participating on equality with white people in any beneficial measures."[6] Brueggemann supports this claim, noting, "The state control of these programs would inevitably lead to discrimination in the South … most domestic workers from SSA provisions reflected open disdain for blacks."[7] Despite Southern prejudice, the federal government still granted states control over these programs, and many Southern states such as Alabama and Georgia used race as a barrier to entry for social security. Barring entry to social security inevitably affected those who were most economically vulnerable, which disproportionately were African Americans. Therefore, the Social Security Act's exclusions illustrate how New Deal legislation exacerbated racial inequalities by systematically denying essential benefits to African Americans in need.

Various New Deal employment programs and policies also significantly exacerbated racial inequalities by discriminating against African American workers. For example, Murphy declares that the Agricultural Adjustment Administration (AAA) "evicted black sharecroppers and tenant farmers off of the land they were cultivating." The Civilian Conservation Corps (CCC) "admitted fewer black men, housed them in segregated dormitories and barred black CCC workers from most administrative positions."[8] These actions effectively decreased and limited employment opportunities for African Americans. Removing African Americans from their land also removed a potential revenue stream and resulted in landowners in the US being predominantly White. In addition, The National Recovery Administration (NRA) implemented regulations that kept Black workers out of key fields. Murphy explains that the NRA's cotton industry hours regulation "excluded the central positions where black male workers labored."[9] When wages were supposed to increase, Brueggemann claims, "a disproportionate number of black workers were laid off."[10]

While the New Deal was supposed to increase employment for American citizens, it had the opposite effect on African Americans. Their working conditions were not improved, and they were fired frequently. Furthermore, Murphy states, "the WPA limited black women's employment opportunities to domestic service training programs and sewing programs."[11]

---

[5] Murphy, "African Americans," 5.

[6] Jesse O. Thomas, "Will the New Deal Be a Square Deal for the Negro?," American Decades Primary Sources 4: 1930-1939 (January 1, 2004): 403, Gale in Context: U.S. History.

[7] Brueggemann, "Racial Considerations," 150.

[8] Murphy, "African Americans," 4.

[9] Ibid.

[10] Brueggemann, "Racial Considerations," 148.

[11] Murphy, "African Americans," 9.

These industries paid lower wages compared to those of white women in other industries. Legislation forced African Americans into lower-wage industries, increasing the racial welfare gap. These discriminatory practices, sanctioned and institutionalized by national legislation, clearly highlight how New Deal programs led to lower employment rates and pay for African American workers, thus deepening racial disparities in the workforce.

While the New Deal was purportedly designed to help U.S. citizens to recover financially after the Great Depression, it ended up hurting African Americans through policies on housing, Social Security, and jobs. Practices like erecting housing projects or discrimination that flourished under state control of Social Security provision contributed to a growing racial wealth gap. Understanding these harmful consequences of the New Deal is crucial, as many current social issues stemming from this gap exist to this day, from disparate levels of social capital to a gap in generational wealth. When drafting legislation, racial equity must be a consideration in order to ensure that all Americans will benefit equally from policies, and that they will not inadvertently worsen existing problems.

Works Cited

Brueggemann, John. "Racial Considerations and Social Policy in the 1930s." Social Science History 26, no. 1 (2002): 139-77. JSTOR.

Lee, Bradford A. "The New Deal Reconsidered." The Wilson Quarterly (1976-) 6, no. 2 (1982): 62-76. JSTOR.

Moreno, Paul. "An Ambivalent Legacy: Black Americans and the Political Economy of the New Deal." The Independent Review 6, no. 4 (2002): 513-39. JSTOR.

Murphy, Mary-Elizabeth B. "African Americans in the Great Depression and New Deal." In Oxford Research Encyclopedia of American History. Last modified November 19, 2020. https://oxfordre.com/americanhistory/view/10.1093/acrefore/9780199329175.001.0001/a crefore-9780199329175-e-632. Oxford Research Encyclopedias.

Thomas, Jesse O. "Will the New Deal Be a Square Deal for the Negro?" American Decades Primary Sources 4: 1930-1939 (January 1, 2004): 401-07. Gale in Context: U.S. History.

Valocchi, Steve. "The Racial Basis of Capitalism and the State, and the Impact of the New Deal on African Americans." Social Problems 41, no. 3 (1994): 347-62. JSTOR.

Why Do We Love Gossip? A Sociological Investigation By Yihe Zhu

Abstract

This review examines the functions and motivations behind negative gossip, specifically focusing on gossip that exposes the immoral actions of an individual within a group. Negative gossip not only imposes unfavorable connotations on the subject of the gossip, but also on the individuals who frequently engage in it. So why, then, is negative gossip more prevalent than positive gossip? This article addresses this question by first exploring the prosocial functions of negative gossip: it helps identify group members who violate shared values and warns others to avoid them. In terms of motivation, the article suggests that group members engage in negative gossip because witnessing immoral behavior challenges their own moral beliefs. Sharing such gossip provides a way to release the frustration caused by this moral conflict.

Introduction

At some point in your life, you have likely encountered gossip. Gossip is the exchange of information about a third party, accompanied by an evaluation, with the subject being unaware of the conversation (Sharlene et al., 2017). What makes gossip particularly interesting, and distinguishes it from a simple transfer of information, is the inclusion of an opinion or judgment. This evaluative nature implies that gossip reflects the gossiper's moral values and provides insight into the target's perceived morality (Holland, 1996). Thus, it is reasonable to conclude that both positive gossip (which enhances the third party's reputation) and negative gossip (which damages it) are, in essence, evaluations of the target's morality or immorality.

This paper specifically examines negative gossip, its connection to moral judgments, and its role in a social context. Not surprisingly, negative gossip carries a less-than-favorable reputation. Research indicates that individuals who frequently spread negative gossip are often disliked and distrusted (Turner et al., 2003). Yet, negative gossip remains twice as prevalent as positive gossip (Robbins & Karan, 2019). This creates a paradox: If spreading negative gossip doesn’t appear to enhance the gossiper’s reputation, why is it still so widespread? This paper seeks to resolve this paradox by exploring the question: Why do people engage in negative gossip that targets others' immorality in social settings?

Discussion

Connection Between Gossip and Morality

To assess the impact of negative gossip and moral judgments in a given social context, it is essential to first understand the broader concept of gossiping. Gossip is a pervasive social activity. Research has found that people spend approximately 80% to 90% of their daily conversations engaging in some form of gossip (Emler, 1994). This significant percentage raises the question of how members of society adopted this behavior in the first place. From an evolutionary perspective, gossiping emerges as a byproduct of humans' need for sociality

(Dunbar, 2004). Social cooperation has been vital to human survival. For instance, group hunting can lead to more successful outcomes than individual efforts. However, effective cooperation relies on equal contributions from all members. To ensure this balance, societies have developed a set of rules—norms—that guide behavior (Campbell, 1975). When a group member benefits from the group's efforts without contributing, they are considered a "free-rider" and violate these norms. In the hunting example, a free-rider would enjoy the group's rewards without participating in the hunt. Gossip, an efficient tool for sharing and disseminating information (Foster, 2004), often serves to identify such norm-violators and alert other group members. This mechanism plays a critical role in upholding social norms and ensuring the continued functioning of a group.

The previous analysis highlighted the existence of norms within social groups, which help maintain cooperation, as well as the role of gossip in attributing moral judgments. This process—assessing actions as right or wrong—is closely tied to the foundation of moral judgments. Morality consists of a set of values that define what is acceptable and unacceptable (APA, 2018, Dictionary of Psychology section). In other words, moral judgments function as norms that distinguish between right and wrong, and good and bad. Ostrom (2000) proposed that a central moral norm regulating the functionality of society is the expectation that individuals act in ways that benefit the common good. Any deviation from this expectation is met with punishment, reinforcing the group's norms (Fehr & Fischbacher, 2004). Referring back to the hunting example, a group member who contributes is viewed as morally good. In contrast, one who benefits from the group's efforts without contributing is seen as morally bad—a norm violator or free-rider. This demonstrates a strong connection between gossip and moral judgments within social groups. Both tools are employed to promote cooperation and establish rules to ensure the functioning of a society. From this, we can draw our first conclusion: Morality forms the foundation for social codes of conduct, and gossip serves as a mechanism for spreading and reinforcing these norms.

Prosocial Functions of Negative Gossip

Having established a connection between gossip and moral norms, this paper now focuses on negative gossip that targets the immorality of a subject. According to the moral foundations theory (Graham et al., 2009), five universal moral foundations—harm/care, fairness/reciprocity, ingroup/loyalty, authority/respect, and purity/sanctity—shape an individual’s perception of morality and their emotional response to violations of those moral principles (Cannon et al., 2010). Gossip is not merely the exchange of information; it also involves an evaluation of that information. This evaluation frequently includes opinions about how the target's behavior is morally unacceptable (Holland, 1996). Studies show that when faced with norm-violating versus non-norm-violating behavior, participants are more likely to gossip about the former (Shallcross et al., 2011). In other words, people are more inclined to gossip about behaviors that are deemed morally wrong. A study conducted in a sorority house supports this finding, revealing that members who exhibited selfish, manipulative, and uncooperative behavior

(all immoral actions) were most often the subjects of gossip (Keltner, 2008). The prosocial gossip theory helps explain this phenomenon. As previously mentioned, gossip is an effective way of spreading and reinforcing norms within a group.

More specifically, gossip that highlights immoral actions—those that violate accepted group norms—serves as a warning to other group members. Despite its negative connotations, negative gossip discussing unacceptable behavior is argued to have prosocial effects. Recipients of such gossip are likely to avoid interacting with the individual known for immoral behavior. Additionally, group members are more likely to behave morally themselves, as they wish to avoid becoming the target of gossip and developing a negative reputation in the future (Sommerfeld et al., 2007; Beersma & Van Kleef, 2011). This, in turn, leads to increased cooperation within the group. For example, in a group where cheating on a spouse is morally unacceptable, when negative gossip spreads about someone who has cheated, other members are made aware of this moral violation and will condemn the individual. Furthermore, group members may avoid cheating themselves, fearing they could become the next target of gossip. As a result, the group's moral values are reinforced, and future norm violations are deterred. While the word "gossip" may carry a negative connotation, negative gossip can have prosocial effects by promoting cooperation and reinforcing group norms through the spread of information and warnings.

Motivations of Negative Gossip: A Paradox

Research suggests that those who spread negative gossip about a target typically do not directly benefit from doing so. In fact, they are often viewed more negatively than those who spread positive gossip (Lian et al., 2022). This creates a paradox: Why do people continue to spread negative gossip if they do not gain any direct advantages? As discussed in earlier sections, gossip that exposes immoral behavior can have prosocial effects on a group. However, this raises the question: Do people engage in negative gossip solely out of a sense of overwhelming altruism, aiming to make a prosocial contribution?

In response to this paradox, some scholars argue that the motivation behind spreading negative gossip is rooted in frustration and the release of that frustration. They suggest that when selfish or exploitative behavior (such as the free-riding behavior mentioned earlier) occurs, it conflicts with an individual's prosocial moral values (e.g., care, loyalty, cooperation). This violation of one's beliefs elicits negative emotions, such as frustration, prompting the individual to take action. By addressing the norm-violating behavior through gossip, the individual experiences relief from their frustration. This, in turn, motivates them to restore cooperation within the group and prevent future violations of norms. In a study by Feinberg, Willer, Stellar, and Keltner, participants were given the opportunity to gossip about someone behaving immorally by being selfish and exploitative. A significant number of participants chose to gossip, sharing evaluative information about the immoral act with another individual involved in the scenario to protect them from exploitation. Physiological measures indicated that participants experienced negative arousal after witnessing the immoral act, but their frustration decreased

once they engaged in gossip, exposing the exploitative individual's behavior (Feinberg et al., 2011). As a result, the negative gossip shared helped protect the individual at risk of being exploited, while the gossipers experienced relief from their frustration caused by witnessing actions that contradicted their moral values. The spreading of negative gossip that exposes immorality is not solely driven by altruistic intentions to benefit the group. Rather, it is primarily motivated by individuals' frustration when they encounter immoral actions that challenge their values and their desire to release these negative emotions.

Ostracism: A Common Concomitant Phenomenon

As previously mentioned, negative gossip has prosocial effects on a group by warning group members and encouraging them to avoid the immoral target. This leads to a particular phenomenon—group ostracism. Researchers have hypothesized that certain levels of social exclusion, when paired with negative gossip, can promote even higher levels of group cooperation than gossip alone (Willer et al., 2011). In a study conducted by Feinberg in 2014, participants were asked to play three types of games: a normal game, a gossip game, and a gossip-with-ostracism game. The general objective of the games was for participants to decide how much money they wanted to allocate to the group versus how much they wanted to keep for themselves. In the gossip game, the choices of each participant were revealed to other group members, who could then send a gossip note to the next round of participants. The gossip-with-ostracism game included an additional feature: at the beginning of each round, group members were allowed to anonymously vote to remove a participant from the group. Supporting the earlier arguments about the prosocial effects of gossip, a comparison between the gossip game and the normal game showed that participants contributed significantly more to the group when gossip was present. However, further comparison with the gossip-with-ostracism game revealed that participants cooperated even more when both gossip and ostracism were at play.

After participants received negative gossip highlighting a norm-violating behavior, they ostracized the individual by voting them out. The study argued that merely being aware of the possibility of being negatively gossiped and ostracized significantly reduced the likelihood of committing actions considered immoral by the group, such as selfishly allocating more money to themselves. Additionally, when the ostracized members of the gossip-with-ostracism game were allowed to return for another round, their contributions to the group significantly increased. This suggests that ostracism, induced by negative gossip, compelled immoral participants to change their behavior and align more with the group's moral values to avoid further ostracization (Feinberg et al., 2014).

In summary, ostracism—a specific social phenomenon—often occurs alongside negative gossip and can lead to a more substantial increase in group cooperation by directly punishing members who fail to conform to the group's values.

Conclusion

This study aimed to explore why individuals engage in negative gossip targeting immorality by examining its functions and effects within a social group. After establishing a connection between gossip and the moral norms within a group, we concluded that negative gossip plays a significant role in promoting prosocial behavior and enhancing group cooperation. Negative gossip, which exposes an individual's violation of a group's moral values, serves as a warning to others, alerting them to distance themselves from the individual. It also reinforces the group's values as members become more aware of the potential consequences of being the subject of gossip.

Furthermore, we discussed how frustration drives individuals to partake in negative gossip when they witness immoral acts that challenge their values. Engaging in gossip that exposes these acts allows the gossiper to release their frustration and feel as though they are reprimanding the behavior. Finally, we analyzed the relationship between gossip and group ostracism. Gossip facilitates ostracism by informing group members about a target's failure to uphold group values, leading to the target's exclusion as a form of punishment. Consequently, the group's cooperation is strengthened as norm violators are removed, and members are motivated to adhere more closely to the group's values to avoid being ostracized.

While this review has highlighted several factors that contribute to understanding the motivations and functions of negative gossip, there are still elements that could complicate our understanding and require further investigation. For instance, as the moral values and norms of a group evolve, would individuals previously targeted by negative gossip and ostracism be forgiven and reintegrated into the community, or would they continue to be shunned as immoral members? Additionally, what if the information spread by the gossiper is entirely malicious and false? Could group members recognize such faults, and would the gossip still yield prosocial effects, as valid negative gossip does?

Despite these unanswered questions, based on the analyses presented, we conclude that negative gossip targeting immoral individuals has prosocial effects on a group by promoting overall cooperation. Furthermore, people engage in negative gossip as a response to frustration when their moral values are challenged, using gossip as a means of releasing such frustration.

Works Cited

APA Dictionary of Psychology. (2018, April 19). American Psychology Association. Retrieved March 5, 2024, from https://dictionary.apa.org/morality

Beersma, B., & Van Kleef, G. A. (2012). Why people gossip: An empirical analysis of social motives, antecedents, and consequences. *Journal of Applied Social Psychology*, *42*(11), 2640-2670. https://doi.org/10.1111/j.1559-1816.2012.00956.x

Campbell, D. T. (1975). On the conflicts between biological and social evolution and between psychology and moral tradition. *American Psychologist, 30*(12), 1103–1126. https://doi.org/10.1037/0003-066X.30.12.1103

Cannon, P. R., Schnall, S., & White, M. (2010). Transgressions and expressions. *Social Psychological and Personality Science*, *2*(3), 325-331. https://doi.org/10.1177/1948550610390525

Dunbar. (2004). Gossip in Evolutionary Perspective. *Review of General Psychology*, *8*(2), 100–110. https://doi.org/10.1037/1089-2680.8.2.100

Emler, N. (1994). Gossip, reputation, and social adaptation. In R. F. Goodman & A. Ben-Ze'ev (Eds.), *Good Gossip* (pp. 117–138). University Press of Kansas.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63-87. https://doi.org/10.1016/s1090-5138(04)00005-4

Feinberg, M., Willer, R., Stellar, J., & Keltner, D. (2012). The virtues of gossip: Reputational information sharing as prosocial behavior. *Journal of Personality and Social Psychology, 102*(5), 1015–1030. https://doi.org/10.1037/a0026650

Feinberg, M., Willer, R., & Schultz, M. (2014). Gossip and ostracism promote cooperation in groups. *Psychological Science*, *25*(3), 656-664. https://doi.org/10.1177/0956797613510184

Fernandes, S., Kapoor, H., & Karandikar, S. (2017). Do we gossip for moral reasons? The intersection of moral foundations and gossip. *Basic and Applied Social Psychology*, *39*(4), 218-230. https://doi.org/10.1080/01973533.2017.1336713

Foster, E. K. (2004). Research on Gossip: Taxonomy, Methods, and Future Directions. *Review of General Psychology*, *8*(2). https://doi.org/10.1037/1089-2680.8.2.78

Haidt, J., Graham, J., & Joseph, C. (2009). Above and below left–right: Ideological narratives and moral foundations. *Psychological Inquiry*, *20*(2-3), pp. 110–119. https://doi.org/10.1080/10478400903028573

Holland, M. G. (1996). What is Wrong with Telling the Truth? An Analysis of Gossip. *JSTOR*, *33*(2), 197-209. https://www.jstor.org/stable/20009858

Keltner, D., Van Kleef, G. A., Chen, S., & Kraus, M. W. (2008). A reciprocal influence model of social power: Emerging principles and lines of inquiry. *Advances in Experimental Social Psychology*, 151-192. https://doi.org/10.1016/s0065-2601(07)00003-2

Lian, H., Li, J. (K.), Pan, J., Du, C., & Zhao, Q. (2023). Are gossipers looked down upon? A norm-based perspective on the relation between gossip and gossiper status. *Journal of Applied Psychology, 108*(6), 905–933.

Ostrom, E. (2000). Collective action and the evolution of social norms. *Journal of Economic Perspectives*, *14*(3), 137–158. https://doi.org/10.1257/jep.14.3.137

Robbins, M. L., & Karan, A. (2019). Who Gossips and How in Everyday Life? *Sage Journals*, *11*(2). https://doi.org/10.1177/1948550619837000

Shallcross, S. L., Howland, M., Bemis, J., Simpson, J. A., & Frazier, P. (2011). Not "capitalizing" on social capitalization interactions: The role of attachment insecurity. *Journal of Family Psychology, 25*(1), 77–85. https://doi.org/10.1037/a0021876

Sommerfeld, R. D., Krambeck, H.-J., Semmann, D., & Milinski, M. (2007). Gossip as an alternative for direct observation in games of indirect reciprocity. *Proceedings of the National Academy of Sciences*, *104*(44), 17435-17440. https://doi.org/10.1073/pnas.0704598104

Turner, M. M., Mazur, M. A., Wendel, N., & Winslow, R. (2003). Relational ruin or social glue? The joint effect of relationship type and gossip valence on liking, trust, and expertise. *Communication Monographs*, *70*(2), 129–141. https://doi.org/10.1080/0363775032000133782

Monetary and Fiscal Policies and the Performance of the US Dollar

By Teodoro Eilert Trevisan

Abstract

The monetary and fiscal policies of the United States and their performance over time have played central roles that are reflected in both today's global economy and their respective historical contexts. To relate these aspects to historical events and the current state of the U.S. economy, an analysis was conducted on the U.S. presidential terms from post-Truman (1953) to the first term of Trump and on the Chairman of the Federal Reserve from Arthur Burns to Janet Yellen, examining their monetary and fiscal indicators as well as the strength of the currency. Furthermore, the research identifies that historically both strong and weak dollar periods have brought benefits and challenges, depending on the global and domestic context, and that the U.S. economy follows identifiable patterns of economic cycles. Therefore, this analysis comprises the relationships of the main political events in American history to the monetary aggregates, price indexes, intrinsic parts of the economic cycles and the dollar's strength choice. Thus, it is clear that the importance of this paper lies on shading light on how the cyclical patterns, aligned with changes in monetary and fiscal policy, reinforce that the strength of the dollar is both a tool and a reflection of the American economic strategy, whose effectiveness depends on adapting to the conditions of each era.

Introduction

While a strong dollar scenario tends to increase imports, reduce exports, and make it harder for American companies to compete on price, a weak dollar scenario, in contrast, tends to increase exports, reduce imports, and make goods and services produced by American companies more attractive to American consumers (Scott Wolla, 2017). Thus, both currency strength levels present pertinent reasons for a country to adopt them—an aspect that will be depicted in this analysis.

Given the existence of this debate, the following sections address changes in monetary policies during the most significant U.S. presidential terms and how these influenced the performance of the U.S. dollar and the monetary aggregates M1, M2, and M3. Each of these reflects, in decreasing order, the liquidity of money and its ability to be converted into consumption.

The monetary and fiscal policies of the United States and their performance over time have played central roles that are reflected in both today's global economy and their respective historical contexts. To relate these aspects to historical events and the current state of the U.S. economy, an analysis was conducted on the U.S. presidential terms from post-Truman (1953) to Trump's first presidency, examining their monetary and fiscal indicators as well as the strength of the currency.

Monetary Policy History

The 1950s were marked by robust economic growth, reflected in an average gross domestic product (GDP) increase of 3.6% annually. This performance was driven by the post-war recovery and rising domestic and international demand. During this period, the Federal Reserve adopted a conservative monetary policy, balancing economic growth with the need to keep inflation under control. As a result, the monetary aggregates experienced more moderate growth rates, below 10% per year. By the end of the decade, Dwight Eisenhower's decision to reinforce the gold standard underscored a commitment to monetary stability, strengthening the dollar and establishing it as a reference for global finance and other currencies, solidifying its position within the Bretton Woods system.

The following decade was characterized by two costly programs: the Apollo mission and the Vietnam War, in addition to the commitment to the Western agreement through Bretton Woods. Consequently, the expected outcome was a rise in M2 and M3 monetary aggregates (above 30% during the two Democratic mandates of the period). This resulted in an imbalance popularly known as the Great Inflation.

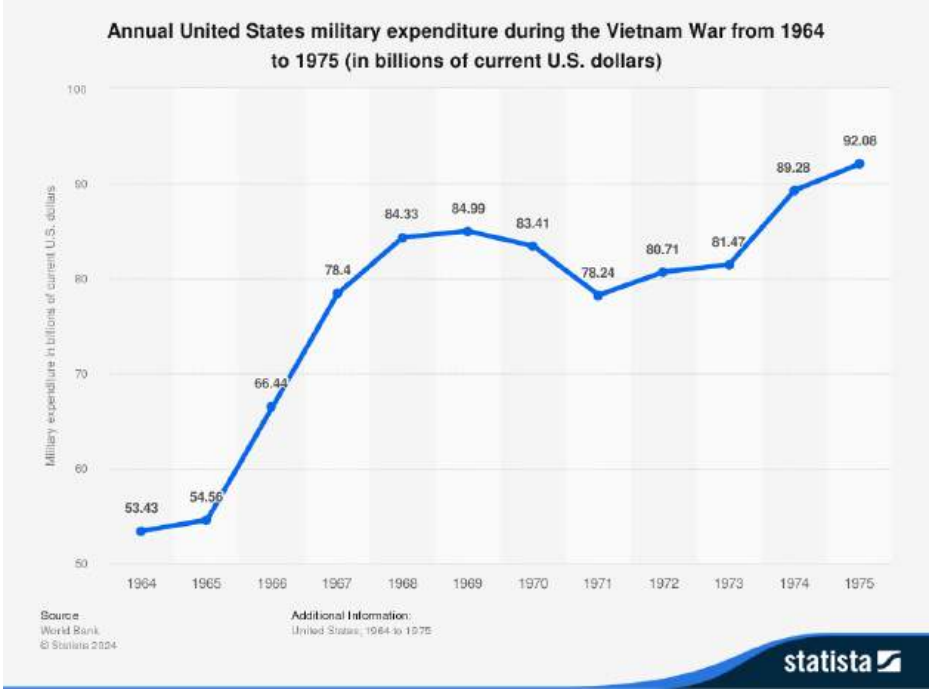


Fig 1: Growth in military expenditure during the Vietnam War, which contributed to the Great Inflation. Source: World Bank @Statista2024

In an attempt to correct the economic balance, reduce inflation, and lower unemployment, Nixon introduced a new economic policy that marked him as the first Republican president to relinquish the strength of the dollar. More specifically, the collapse of the Bretton Woods system and the end of the gold standard in 1971 marked a turning point: the Fed began aggressively expanding the money supply to also address the oil crisis and stagflation,

reflecting fiscal problems and directly impacting the monetary aggregates, particularly M2 and M3 (both increased by 42%).

By the late 1970s, during Jimmy Carter's presidency (1977–1981), the U.S. continued to struggle with inflation, and monetary policy became even more restrictive. The Federal Reserve adopted high interest rates to control double-digit inflation, significantly reducing the money supply (further impacted by the second oil shock).

The high-interest-rate strategy persisted, but inflation control was limited, reflecting the complexity of managing the economy in a period of crisis. Rising interest rates increased the cost of credit, directly affecting economic growth and demand.

During Ronald Reagan's presidency (1981–1989), U.S. monetary policy became more focused on fighting inflation, with support from then-Federal Reserve Chairman Paul Volcker. Reagan combined an expansionary fiscal policy—marked by significant tax cuts and increased military spending (Reaganomics)—with a contractionary monetary policy, implementing high interest rates (reaching 20%). The early 1980s recession resulted in short-term economic deceleration but led to substantial long-term recovery.

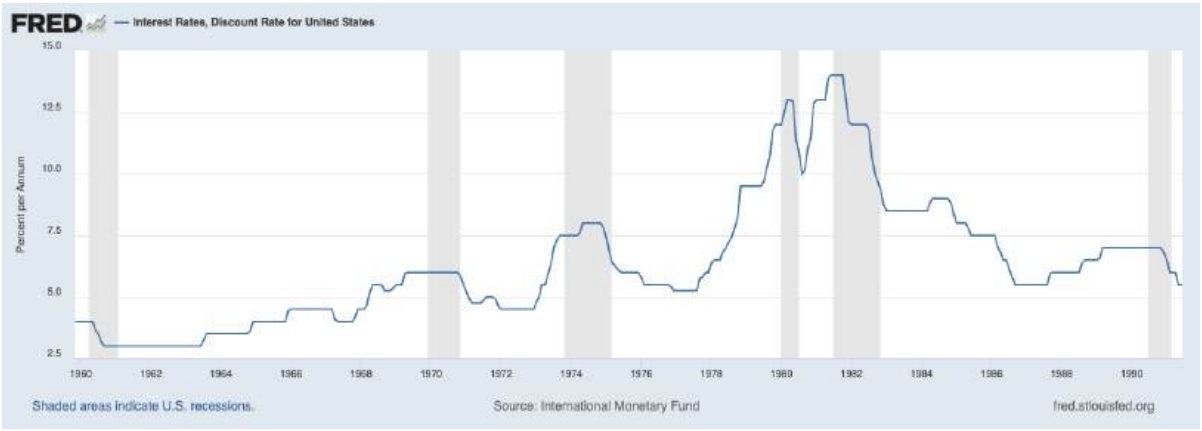


Fig 2: High interest rates as part of Reaganomics, marking the contractionary monetary policy side of it. Source: International Monetary Fund @Fred St. Louis

During Bill Clinton's presidency (1993–2001), the U.S. experienced one of the longest economic expansions in history, with the largest fiscal surpluses in U.S. history ($236 billion in 2000). Monetary policy was characterized by a more flexible approach, with the Federal Reserve using moderate interest rates to maintain economic stability while allowing growth. The increase in M2 and M3 reserves (which grew by almost 30% from the first to the second term) was a direct consequence of this phase of robust economic growth.

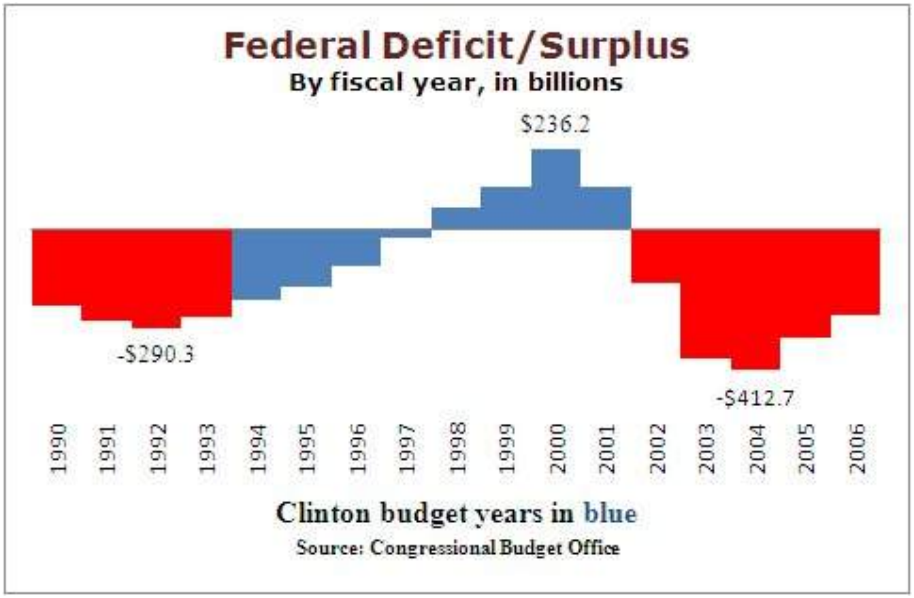


Fig 3: Largest fiscal surpluses in U.S. history during Bill Clinton's presidency. Source: Congressional Budget Office.

The two terms of George W. Bush (2001–2009) were also marked by significant economic expansion, evidenced by an average GDP growth of 2.6% per year between 2001 and 2007. However, this period also saw a brief recession in 2001 and a tumultuous end, with the 2007–2008 financial crisis and the rise of China. The Federal Reserve adopted quantitative easing and very low interest rates to help contain the crisis's effects. Consequently, there was a significant increase in the money supply (M2 and M3 again approaching 30% growth over both terms), along with the Fed injecting liquidity into the financial system, accelerating M1 growth (24% and 15% in the respective terms). During this period, the dollar's performance was impacted by financial uncertainty, and economic recovery was gradual.

Barack Obama (2009–2017) continued to implement economic stimulus measures, with the Fed maintaining low interest rates and employing quantitative easing, a measure that, although unconventional, became more attractive to central banks due to the ineffectiveness of traditional measures like interest rate adjustments (Bernanke, 2017). During Obama's presidency, the U.S. economy began recovering from the 2008 Great Recession, and the dollar showed signs of strengthening.

Finally, Trump's first administration (2017–2021) sought, in general terms, to normalize monetary policy, with attempts to raise interest rates. However, the COVID-19 pandemic, which emerged at the end of his term, forced a reversal of this strategy, with a new round of fiscal and monetary stimulus leading to an extraordinary increase in M1 during this period (432%), along with impacts on M2 and M3 (45%).

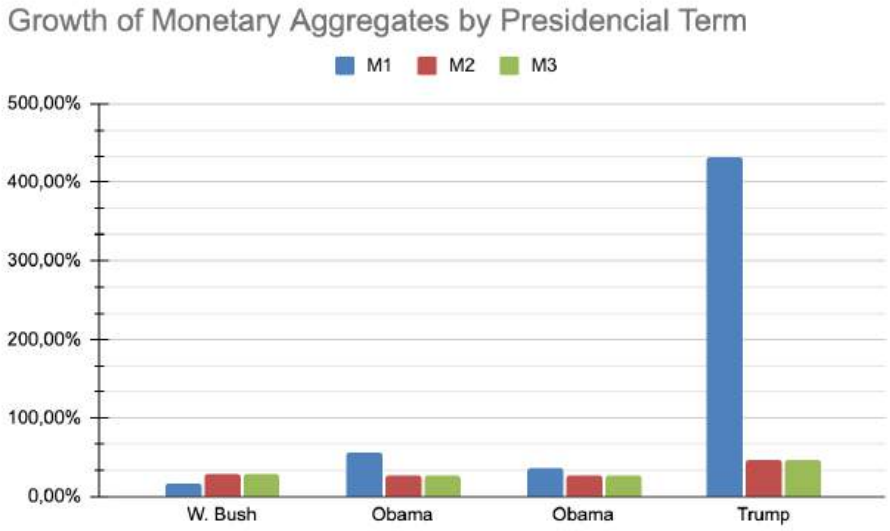


Fig 4: Extraordinary growth in monetary aggregates during Trump's first term. Source: the authors (2024). Prepared from the Federal Reserve of St. Louis.

In summary, M1, which includes cash in circulation and demand deposits, has a direct relationship with whether monetary policy is restrictive or expansionary. During periods of high interest rates, such as under Volcker's administration, M1 tends to grow more slowly, whereas during periods of monetary stimulus, as seen under Clinton and Obama, it accelerated.

M2, which includes M1 plus savings deposits and other liquid assets, and M3, the broadest indicator (encompassing M2 plus large deposits and investment funds), follow the same trend, growing faster during expansionary policy periods. Additionally, both aggregates experienced substantial expansion during financial crises and quantitative easing periods, especially during the Bush and Obama administrations. During crises, the Fed needed to increase the monetary base to ensure liquidity in the financial system.

Regarding the performance of the dollar, it is intrinsically linked to the expansionary or contractionary monetary policies adopted by presidents and Fed officials. For example, under Nixon, the abandonment of the gold standard and the implementation of expansionary monetary policies led to the dollar's devaluation as the money supply was increased to stimulate the economy. Conversely, during the Carter and Reagan administrations, Volcker led a contractionary monetary policy with high interest rates to combat inflation. This resulted in a strong appreciation of the dollar, particularly in the early 1980s, affecting American exports due to the loss of competitiveness.

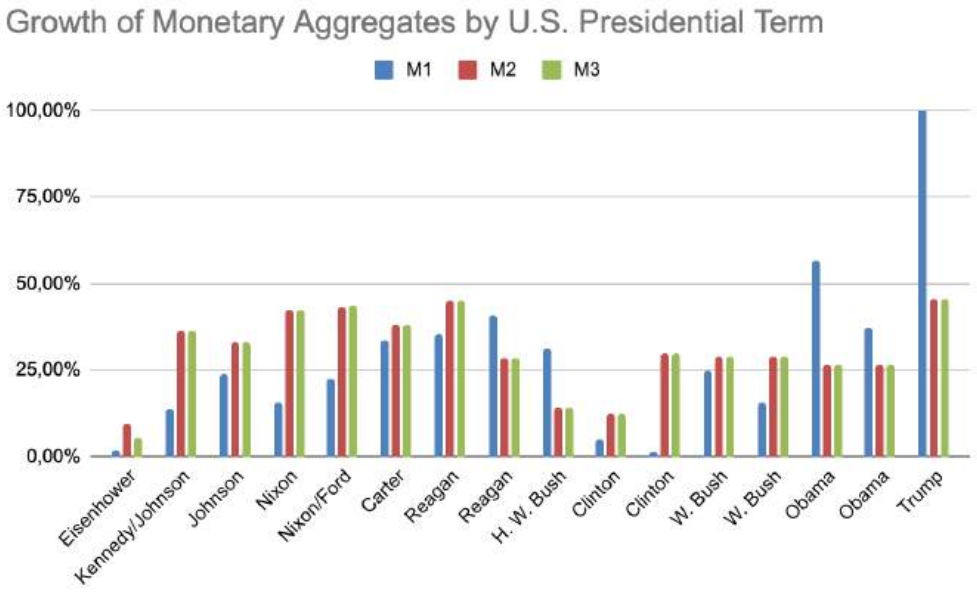


Fig 5: General overview of growth in monetary aggregates by USA presidential term. Source: the authors (2024). Prepared from the Federal Reserve of St. Louis.

Impact of Monetary Policy - Analysis of the Relationship Between M2 and M3 Monetary Aggregates and Producer Price Index (PPI)

In the U.S. economy, where a significant portion of the population has access to investments and more sophisticated financial instruments, M2 and M3 aggregates are more relevant for understanding the impacts of monetary policy on inflation measured by the Producer Price Index (PPI). This economic indicator measures price variations that producers receive for their goods and services before reaching the final consumer and precedes consumer inflation (CPI), as cost increases for producers are eventually passed on to consumers.

During Eisenhower's two terms in the 1950s, the moderate growth of M2 and M3 aggregates, reflecting a conservative monetary policy, resulted in low influence on the PPI (which grew an average of 4% per term), demonstrating stability in production costs.

However, the economic burdens of the Apollo mission and the Vietnam War, along with compliance with the Bretton Woods agreement, placed immense strain on the American economy. Detailing how this led to the Great Inflation, the structural problem highlighted by Fed History (2013) as the "Triffin Dilemma" emerged. This occurs when a country issues a global reserve currency as a medium of exchange due to its international importance (in this case, the United States). The source explains that the stability of this currency is questioned when the country persistently runs current account deficits to supply this reserve. As a result, Nixon's presidency (1969–1974) began with the U.S. facing one of the worst inflation crises in modern history.

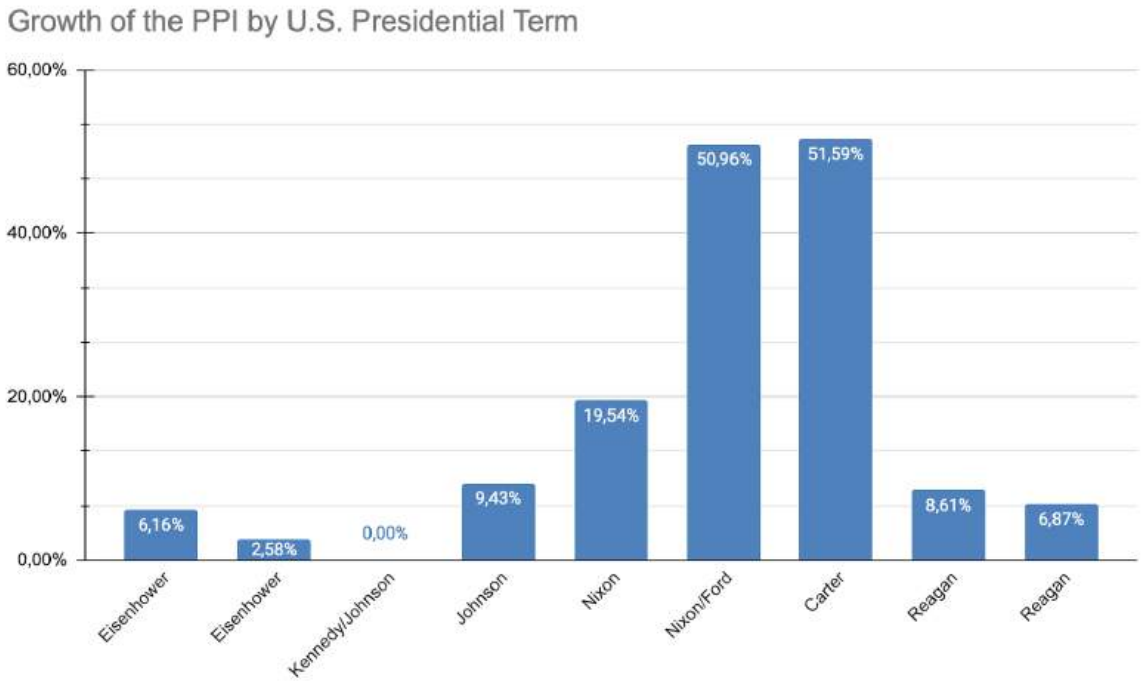


Fig 6: Increased growth of PPI during the Great Inflation. Source: the authors (2024). Prepared from the Federal Reserve of St. Louis.

During the Nixon administration (1969–1974) and the new economic policy implemented by Fed Chairman Arthur Burns, the growth in M2 and M3 aggregates contributed to a significant increase in the PPI (19%). The double-digit inflation caused by overly expansionary policies, which was also Burns' first major error (Pierce, 1979), contributed to the onset of stagflation, exacerbated by supply shocks such as the oil embargo. The continued growth of these aggregates reinforced inflationary pressures, raising the PPI to 50% during Nixon's second term and Ford's presidency, highlighting how excess liquidity in the economy, combined with the energy crisis, drove up production costs. Similarly, during the stagflation period, the increase in aggregates during Carter's term led to an inflationary surge in the PPI, again surpassing 50%. This illustrates the failure of monetary policies at the time, or in other words, the defeat of Burns' monetary policy flexibility in dealing with supply shocks and rising global costs (Pierce, 1979).

In the 1980s, under Reagan, the Federal Reserve adopted contractionary policies despite significant aggregate growth early in his administration. The high-interest-rate policy under Paul Volcker curbed cost inflation, reducing PPI growth from 51% to 8%. This period demonstrates the impact of monetary control in reducing cost pressures in the production chain. Thus, economic recovery and dollar stabilization also reflected the success of Volcker's monetary policy, which managed to control inflation and stabilize the currency while restoring confidence in the U.S. monetary system.

During the 1990s, under Clinton, the well-managed growth of M2 and M3, along with notable fiscal surpluses, resulted in a relatively stable PPI (ranging from 10% to 8% across Clinton's terms). This period was characterized by a balance between fiscal and monetary policy, ensuring controlled inflation despite economic expansion. Moreover, monetary stability, with controlled inflation, helped consolidate the dollar's strength relative to other currencies in the international market.

Under George W. Bush (2001–2009), the accelerated growth of M2 and M3, reflecting the monetary response to the post-9/11 recession and the 2008 financial crisis, kept the PPI stable in the first four years but increased by 13% in the second term, demonstrating the effects of global crises and expansionary policies. Additionally, in this context, the implementation of Quantitative Easing (QE) under Obama, which significantly expanded the monetary base and M2 and M3 aggregates, caused the PPI to experience a slight decline (5.83%), reflecting the global recession and disrupted supply chains. However, it also demonstrated the success of this measure in avoiding significant inflationary pressures during a crisis. It is worth noting that QE was applied in such a way that, as Caldentey (2017) explains, the additional liquidity was directed toward financial markets rather than directly impacting the productive economy, meaning producer prices remained stable.

Later, during Trump's first term, the accelerated growth of M3 was accompanied by a moderate PPI change (7%). Expansionary fiscal policies, such as tax cuts, helped mitigate short-term inflationary pressures but increased the risks of long-term fiscal imbalances (Celebi and Welfens, 2020).

In summary, the analysis of the PPI with M2 and M3 aggregates demonstrates that excess liquidity in the economy tends to increase production costs, especially during periods of external shocks. However, contractionary policies like those of the 1980s proved effective in containing the PPI, albeit at the cost of slower economic growth.

On the other hand, in administrations such as Clinton and Obama, the relationship between monetary aggregates and the PPI was more moderate, indicating a more stable monetary policy or one detached from significant inflationary pressures in the productive sector. These patterns reinforce how different economic approaches shaped the relationship between economic liquidity and prices over the decades.

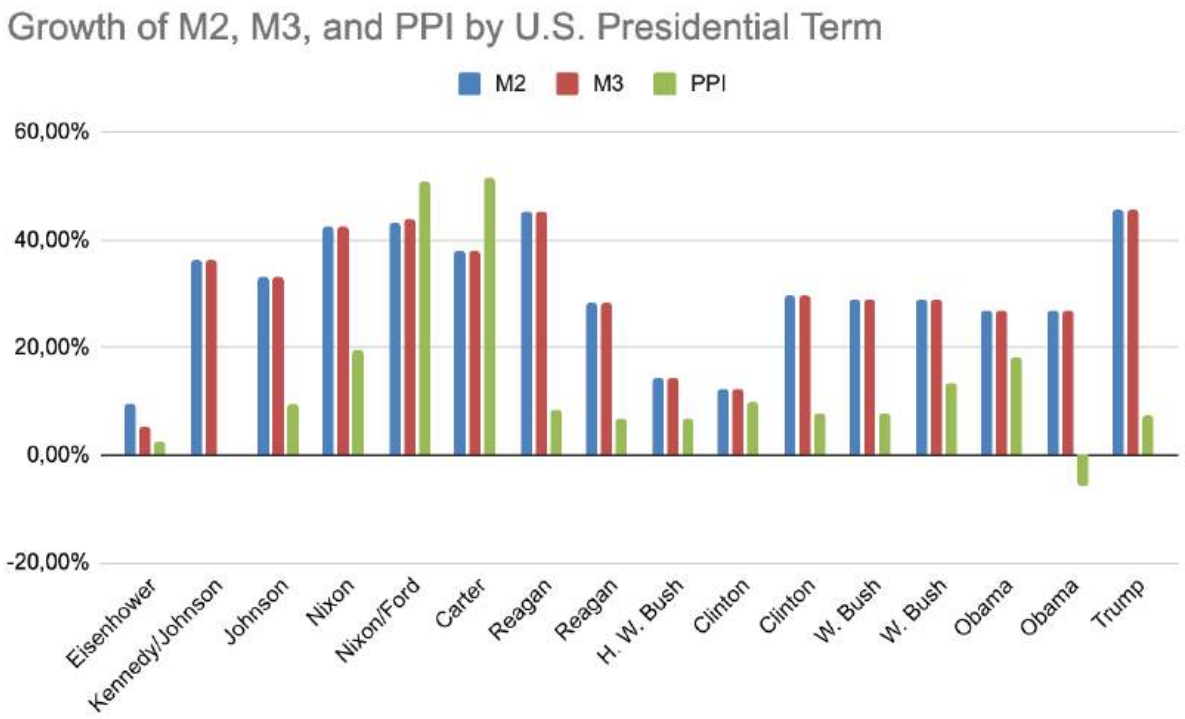


Fig 7: General overview of growth in monetary aggregates compared to PPI growth by USA presidential term. Source: the authors (2024). Prepared from the Federal Reserve of St. Louis.

Analysis of the Relationship Between M2 and M3 Monetary Aggregates and Consumer Price Index (CPI)

Considering the Fed's primary goal of maintaining price stability and full employment, with an inflation target set at 2% per year, the relationship between M2 and M3 monetary aggregates and the CPI during U.S. presidential terms reveals significant economic dynamics. Since the CPI, which measures changes in the price level of goods and services over time, is directly influenced by changes in the money supply, this analysis explores the impact of monetary policies across different periods using historical data to examine correlations between aggregate growth and CPI behavior.

During the Eisenhower administration (8%) and the Kennedy and Johnson administrations (1%), inflation was controlled, reflecting periods of stable economic growth and monetary policies focused on price stability. The lower CPI variation under Kennedy and Johnson is associated with a more gradual expansion of M2 and M3, with minimal inflationary pressure, reinforcing the positive impact of a balanced money supply.

Inflation began to escalate at the end of Lyndon Johnson's term (14%) and increased significantly during the Nixon (20%) and Ford (17%) administrations. The rise in CPI during these periods mirrors the increase in PPI caused by supply crises and insufficiently conservative monetary policies to contain inflation resulting from high public spending (Pearce, 1979), as well as the rapid expansion of M2 and M3 aggregates.

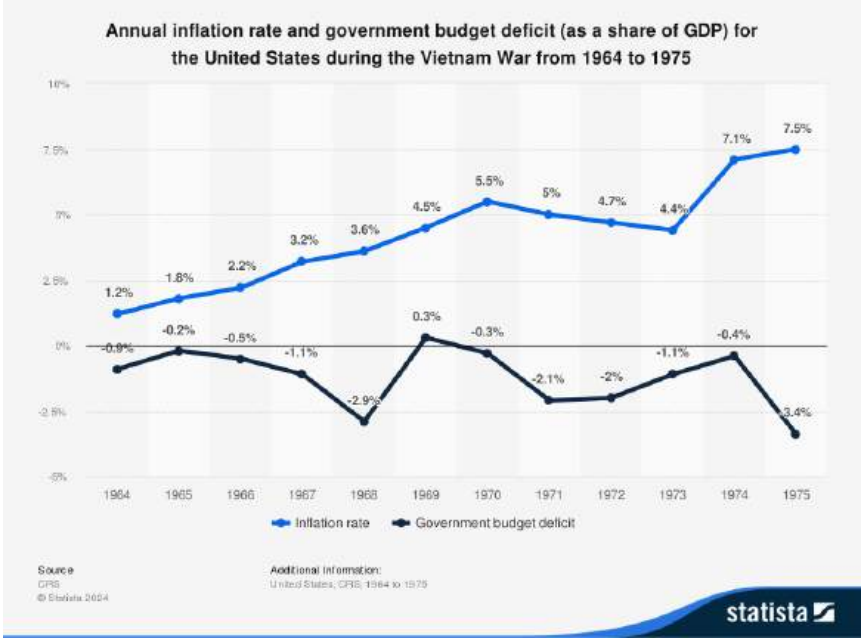


Fig 8: Increased growth in the annual inflation rate for the USA during the Vietnam War. Source: CRS @Statista2024

The inflation crisis reached its peak during Jimmy Carter's presidency (49%). This period, marked by one of the highest CPI growth rates, underscores the impact of expansionary monetary and fiscal policies combined with the second oil shock. Despite attempts to restrict the money supply toward the end of his term, the uncontrolled growth of M2 and M3 contributed to persistent inflation, culminating in stagflation.

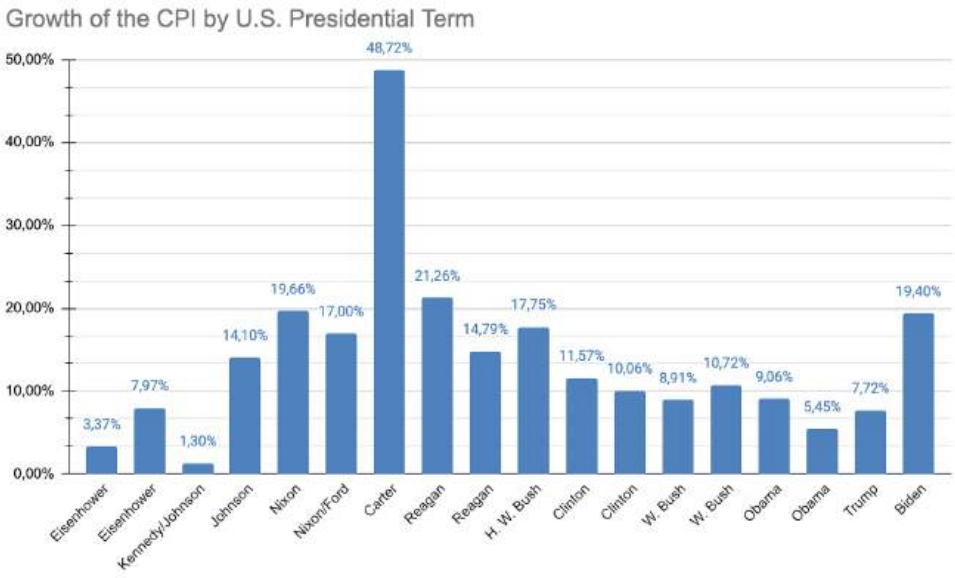


Fig 9: General overview of growth in CPI by USA presidential term. Source: the authors (2024). Prepared from the Federal Reserve of St. Louis.

Under Ronald Reagan (21% and 15%), CPI recorded a decline compared to Carter's term, indicating the partial success of restrictive monetary policies initiated by the Federal Reserve. The slowdown in M2 and M3 growth was crucial in reducing inflationary pressure. During Reagan's second term, economic austerity policies stabilized inflation, demonstrating how adjustments in the money supply can reverse inflationary crises.

The 1990s, under George H. W. Bush (18%) and Bill Clinton (12% and 10%), saw moderate and stable inflation. This period coincided with a controlled expansion of M2 and M3, supported by economic growth and deficit reductions. The monetary policies of this era ensured a positive relationship between economic growth and price stability.

In the 2000s, the terms of George W. Bush (9% and 11%) and Barack Obama (9% and 6%) kept inflation under control, despite economic pressures such as the 2008 financial crisis. During these presidencies, the Federal Reserve adopted a quantitative easing approach, increasing M2 and M3 to stimulate the economy without triggering significant inflation.

Finally, Donald Trump's administration (8%) also reflected moderate inflation despite robust economic growth. However, the increase in CPI toward the end of his term suggests the beginning of inflationary pressures exacerbated by the COVID-19 pandemic.

These analyses highlight that controlling monetary aggregates such as M2 and M3 is essential for managing inflation. Periods of uncontrolled expansion in the money supply are directly associated with CPI spikes, while balanced administration helps maintain economic stability. Even so, while the money supply plays a crucial role in inflation dynamics, its impact is influenced by other factors such as supply shocks, fiscal policies, and global crises. The study of presidential terms reveals that, although monetary expansions often drive inflation, the magnitude of their impact depends on the economic context, the speed of monetary responses, and the government's control measures.

The analysis of the behavior of U.S. monetary and fiscal policy, along with its impacts on the U.S. dollar and economic indices such as M1, M2, M3, CPI, and PPI, reveals evident cyclical patterns that reflect the context of each historical period, highlighting its challenges and decisions. Based on the collected data, it is possible to identify specific dynamics in different economic eras, each shaped by global events, internal crises, and shifts in economic policy. As a result, the following section divides the economic cycles and monetary and fiscal patterns of the United States from 1950 until the first Trump administration.

1950–1960: Post-War American Ascendancy and the Unbacked Dollar

This period was characterized by relative stability in the growth of monetary aggregates and PPI and CPI, supported by policies emphasizing post-war stability. Inflation remained controlled, particularly under the Eisenhower, Kennedy, and Johnson administrations, which exhibited low inflation variations due to the moderate expansion of M2. The unbacked dollar maintained relative stability, facilitating internal economic balance in a world still recovering from World War II.

1960–1972: The Gold-Backed Dollar and Rising Fiscal Deficits

The rising fiscal deficit and financial demands from the Vietnam War and the Apollo program led to an inflationary surge. The initial decline of the Bretton Woods system, coupled with monetary expansion, generated pressures on PPI and CPI that intensified towards the end of this period. With the end of the gold standard, the dollar oscillated between strength and weakness as crises such as the oil shocks and inflation control policies shaped the economy. The dollar began to feel the pressures of increasing fiscal deficits, but its backing in gold maintained its international credibility until 1971 when it was abandoned in favor of a floating system.

1972–1992: Navigating Stagflation

With the abandonment of the gold standard in 1971, the U.S. faced oil shocks and supply crises that led to inflationary surges, peaking during the Carter administration. During Reagan's terms, the Federal Reserve adopted contractionary monetary policies, which were essential in slowing PPI and CPI growth and stabilizing the economy, marking the beginning of a recovery period.

1992–2020: Crises Amidst Prosperity

While the 1990s brought high economic growth and controlled inflation under Bush and Clinton, the new millennium introduced significant economic crises that accompanied the broader economic expansion, namely the 2001 and 2008 crises, which prompted the implementation of Quantitative Easing measures. Thus, the economic stability of the 1990s strengthened the dollar, while subsequent crises tested its resilience. Despite the increase in M2, inflation remained moderate due to subdued demand and market adjustments. However, the explosion in monetary growth during Trump's term signaled new inflationary pressures.

General Analysis

Additionally, based on the analysis of these cycles, several conclusions can be drawn. Among them, notable patterns have emerged in the performance of the U.S. economy.

Firstly, the behavior of CPI and PPI when compared with monetary aggregates highlights the strong relationship between monetary growth and price variations. Periods of high inflation, such as in the 1970s and during the Biden administration (PPI 23% and CPI 19%), coincided with accelerated growth in M2 and M3. On the other hand, periods of stability, such as under Bush and Clinton, reflect monetary supply control.

Furthermore, external shocks, such as the oil crises of the 1970s and the 2008 financial crisis, have had lasting impacts on U.S. economic policies. The stagflation of the Carter era and the quantitative easing responses of the 2000s are clear examples of how global crises have shaped monetary and fiscal management.

Finally, the actions of the Federal Reserve in the 1980s, led by Paul Volcker, also demonstrate that contractionary monetary policies, when strategically implemented, can contain inflationary crises and stabilize the economy in the long run, thus playing a crucial role.

Final Considerations

The historical analysis clearly shows that controlling monetary and fiscal policy is a fundamental element of the economic stability of the United States. However, historical cycles also demonstrate that these controls are deeply intertwined with the global context and the crises of each era. These patterns not only explain the evolution of the U.S. dollar and its economic impacts but also offer valuable lessons for future policymakers, especially in times of economic volatility.

The dollar strength exerts profound effects on the international competitiveness of the United States, influencing the trade balance and attracting investments (Scott Wolla, 2017). The analysis highlights that both strong and weak dollar periods have brought distinct benefits and challenges, depending on the global and domestic context. Ultimately, the US economy follows identifiable cyclical patterns of expansion, peak, recession, and recovery. These cyclical patterns, aligned with shifts in monetary and fiscal policy, reinforce that the strength of the dollar is both a tool and a reflection of American economic strategy, whose effectiveness depends on adapting to the conditions of each era.

Works Cited

Wolla, Scott A. "Is a Strong Dollar Better than a Weak Dollar?" Federal Reserve Bank of St.Louis, 1 Mar. 2015. Accessed January, 2025. https://www.stlouisfed.org/publications/page-one-economics/2015/03/01/is-a-strong-dollar-better-than-a-weak-dollar.

Yellen, Janet L. "Inflation, Uncertainty, and Monetary Policy." Business Economics, vol. 52, no. 4, 2017, pp. 194-207.

Bernanke, Ben S. "Monetary Policy in a New Era." Yale Program on Financial Stability, 2017. Accessed January, 2025. https://elischolar.library.yale.edu/ypfs-documents/9530.

Bligh, Michelle C., and Gregory D. Hess. "The Power of Leading Subtly: Alan Greenspan, Rhetorical Leadership, and Monetary Policy." The Leadership Quarterly, vol. 18, no. 1, 2007, pp. 87-104.

Pierce, James L. "The Political Economy of Arthur Burns." The Journal of Finance, vol. 34, no. 2, 1979, pp. 485-496.

Caldentey, Esteban Pérez. "Quantitative Easing (QE), Changes in Global Liquidity, and Financial Instability." JSTOR, 2017. Accessed January, 2025. https://www.jstor.org/stable/48539963.

Fed History. "Nixon Ends Convertibility of U.S. Dollars to Gold and Announces Wage/Price Controls." Federal Reserve History, 2013. Accessed January, 2025. https://www.federalreservehistory.org/essays/gold-convertibility-ends.

Welfens, Paul J. J., and Korbinian Celebi. "The Economic Impact of Trump: Conclusions from an Impact Evaluation Analysis." RePEc, 2020. Accessed January, 2025. https://ideas.repec.org/p/bwu/eiiwdp/disbei281.html.

Federal Reserve Bank of St. Louis. "M1 (WM1NS)." Accessed January, 2025. https://fred.stlouisfed.org/series/WM1NS.

Federal Reserve Bank of St. Louis. "M2 (WM2NS)." Accessed January, 2025. https://fred.stlouisfed.org/series/WM2NS.

Federal Reserve Bank of St. Louis. "Monetary Aggregates and Their Components: Broad Money and Components: M3 for United States (MABMM301USM189S)." Accessed January, 2025. https://fred.stlouisfed.org/series/MABMM301USM189S.

Federal Reserve Bank of St. Louis. "Producer Price Index by Commodity: All Commodities (PPIACO)." Accessed January, 2025. https://fred.stlouisfed.org/series/PPIACO.

Federal Reserve Bank of St. Louis. "Consumer Price Index: All Items: Total for United States (CPALTT01USM657N)." Accessed January, 2025. https://fred.stlouisfed.org/series/CPALTT01USM657N.

Federal Reserve Bank of St. Louis. "Federal Government Budget Surplus or Deficit (-) (M318501Q027NBEA)." Accessed January, 2025. https://fred.stlouisfed.org/series/M318501Q027NBEA.

Federal Reserve Bank of St. Louis. "Interest Rates, Discount Rate for United States (INTDSRUSM193N)." Accessed January, 2025. https://fred.stlouisfed.org/series/INTDSRUSM193N.

Federal Reserve Bank of St. Louis. "Trade-Weighted Exchange Value of U.S. Dollar vs G-10 Countries (DISCONTINUED) (TWEXMTHY)." Accessed January, 2025. https://fred.stlouisfed.org/series/TWEXMTHY.

Federal Reserve Bank of St. Louis. "Real Broad Dollar Index (RTWEXBGS)." Accessed January, 2025. https://fred.stlouisfed.org/series/RTWEXBGS.

Artificial Intelligence in Ultrasound: Unlocking New Possibilities in Imaging
By Deniz Oktar

Abstract

Ultrasound (US) is one of the most commonly used imaging tools, especially for the abdomen, superficial organs and vascular structures. Main advantages of US examination include its non-ionizing nature, relatively low cost, non-invasiveness, portability, and real-time imaging. However, the US has some drawbacks arising from characteristics of US imaging, such as significant operator dependency, subjectivity in image interpretation, relatively low resolution, and limited penetration. Artificial intelligence (AI) technology has the potential to provide more efficient evaluation of US images which may help to overcome some of these drawbacks enabling more repeatable outcomes in US examination. In this paper, we review the use of AI in diagnostic US.

Introduction

Diagnostic ultrasound (US) is a non-invasive imaging tool for evaluating the internal organs of the body. US images are created by sending US pulses into tissues using a "probe". US probes, also called transducers, produce sound waves that have frequencies above the upper threshold of human hearing that is 20,000 Hz. The US waves are reflected from the tissues in varying amounts depending on reflection properties. The returning signals are then collected by the US probe and displayed as an image. Most transducers in current use operate at much higher frequencies in the range of 2-20 megahertz (MHz). Since there is no ionizing radiation, US imaging is generally considered as safe when used correctly (1, 2). Compared to other medical imaging modalities like conventional roentgenograms, computed tomography (CT) or magnetic resonance imaging (MRI), US has some advantages like absence of ionizing radiation, safety, relatively low cost, wide availability, and real-time imaging (3).

Most common application areas of the US include diagnosis of abdominal pathologies, pregnancy monitoring, thyroid screening, vascular examinations, or breast cancer detection. These examinations are typically performed by highly trained personnel specialized in these fields.

The most recent advancement in diagnostic US is the incorporation of AI to this method. According to a recent study, published in Journal of the American College of Radiology, AI in the US can significantly enhance image quality, improve image acquisition, aid in disease detection, and optimize clinical workflows (4). The tremendous potential of combining clinical expertise with computational technologies has made the computer-assisted US a popular research field. When we search the terms "Ultrasound" and "Deep Learning" on Web of Science we observed that there is an exponential increase in the number of academic publications, reflecting the growing interest in this emerging technology in various research areas (Figure 1). In this paper, application of AI in diagnostic US is briefly reviewed.

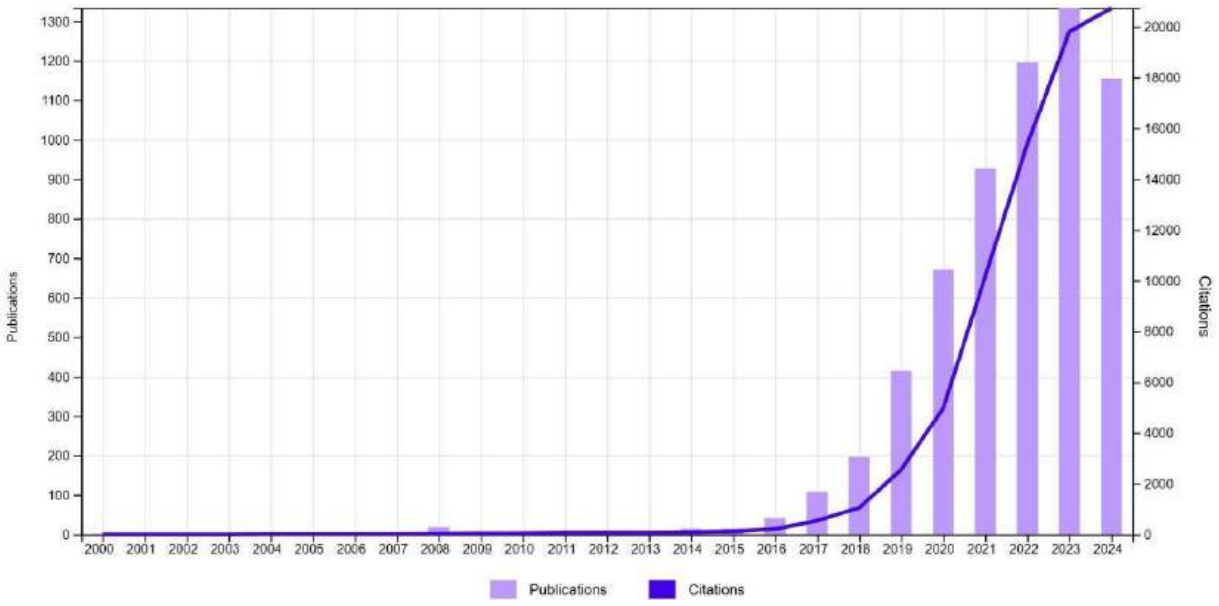


Figure 1. Exponential growth in the number of articles and citations indexed in Web of Science with the when the search terms "Ultrasound" and "Deep Learning" applied, reflects the growing interest in this technology.

Challenges in diagnostic US

The US is the first line and sometimes the only imaging tool in many medical conditions. With the recent developments in US technologies, a hand carried wireless US transducer connected to a mobile phone has the capabilities of what a conventional US unit has such as displaying images. The US now holds the potential to become the new stethoscope with its well-known advantages and portability. On the other hand, there are some drawbacks associated with US imaging. First and foremost, it is operator dependent and it may be challenging to obtain reliable and similar results across different users. The US requires long and specialized training. Another drawback is that the image interpretation depends on experience and skills of the operator, leading to subjectivity. The resolution of US images is less than other imaging tools making them harder to read compared to X-ray, CT and MRI. It may be difficult to accurately diagnose some conditions, especially for the structures located deep within the body. Limited penetration depth, limited quality of images, inability to evaluate bone and to observe air-filled areas are among disadvantages of this modality. Although technological advancements help to address some of these challenges, there remain considerations in the use of the US. AI technology offers the potential to provide more efficient and objective evaluation of US images, also providing quantitative assessment (5).

Overview of artificial intelligence

Artificial intelligence (AI), is a term that encompasses subfields within it, such as, machine learning (ML), deep learning (DL) and convolutional neural networks (CNNs). Figure 2 represents the hierarchical relationships between these branches.

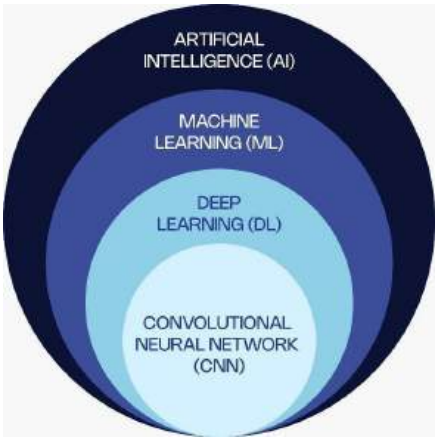


Figure 2. Simplified hierarchical relationships between AI, ML, DL and CNN.

All inclusive, AI is defined as `the use of computer programs that have some of the qualities of the human mind, such as the ability to understand language, recognize pictures, and learn from experience` (6). On simpler terms, AI, although much broader, is like an imitation of the human brain that can process weighted values, analyze them, and predict without human intervention. Therefore, radiological imaging is a field that can benefit from AI algorithms by training them to visualize and detect abnormalities within our biological premises. **Machine Learning (ML)** is a branch of AI. ML is defined as `the process of computers improving their own ability to carry out tasks by analyzing new data, without a human needing to give instructions in the form of a program` (6). This process includes training, testing and validation steps. The predictions in ML can be generated through *supervised learning*, where algorithms learn patterns from existing data, or *unsupervised learning*, where algorithms discover general patterns in data automatically. The most commonly used ML method in US imaging is the supervised learning. In this method, data sets that are labeled by experts are called 'ground truth'. The term 'ground truth' refers to correct or true labels for a given data set. This way ML models can be trained to examine US images and look for certain appearances of diseases such as cancer (7, 8).

Deep learning (DL), is a subfield of ML. DL can be referred to as a replica of a human brain that's learning. DL autonomously extracts a complex hierarchy of features from images, unlike classical ML algorithms which rely on hand-coded sets of instructions for decision (4). It is based on multiple layers of artificial neural networks (ANNs) which is a system mimicking the way neurons operate in the human brain.

ANN consists of interconnected nodes, or neurons organized in multiple layers. They include three main types of layers. *Input layer* is the layer receiving initial data and transmits them to subsequent layers for processing. *Hidden layers* perform complex computations on input data extracting meaningful features and patterns and pass the results to neurons in the next layer. This process continues until the final output is generated. *Output layer* is the final layer producing the network's predictions or outputs based on information from hidden layers.

DL was used in many parts of medicine for segmentation of images produced via devices used. Unlike conventional ML algorithms, DL is capable of learning from enormous amounts of data, and able to produce remarkably accurate findings. Language translation, image recognition, and personalized medicines are some examples of deep learning applications (4, 9).

Convolutional neural networks (CNNs) are subcategories of ANNs, whose input type is explicitly assumed to be images. CNNs learn in a way very similar to human beings. Babies are born without recognizing objects like birds or dogs. As we grow, we gradually learn to associate specific shapes and colors with particular items. We gain the ability to distinguish between animals by recognizing defining features such as tails in dogs or beaks in birds. Very simply, the CNNs operate same way. By analyzing labeled training images, CNNs can identify key features that help them differentiate between objects in those images. Architecture of CNNs is extremely complex, some systems having more than 100 layers, meaning billions of connections among neurons. CNNs use multilayer algorithms to classify visual inputs through multiple convolutional layers that are either fully connected or pooled. These networks have been demonstrated to be efficient in performing medical imaging tasks and have the capability to learn different features of medical imaging (4, 8, 10, 11).

Main applications of AI in ultrasound imaging

There are few main research areas where AI solutions can enhance US practice. First of all, ML programs can be trained to examine images and look for certain indicators of disease. This way algorithms can help reduce the long and difficult learning curve needed for US scanning and increase diagnostic accuracy of the examiners. Developing a system that reduces the level of expertise needed for US could have significant impact in health system. With the help of such a system, any person in a remote area with a basic anatomical knowledge would be able to do an US exam and send the images to an expert for evaluation regardless of physician`s location. This is also helpful in point of care (bed-side) US, which is often performed by less-skilled clinicians or technicians. In recent years AI algorithms have demonstrated good performance in diagnosing various diseases from medical images, sometimes even surpassing human accuracy. Most of these studies focus on CT or MRI imaging. However, considerable amount of research has also been conducted on applying AI to US imaging. Its impact on developing or less developed countries can be even more profound. For instance, since there are not enough experienced physicians or sonographers in these countries, many women do not

receive US follow up during their pregnancy, which is necessary for the well-being of both mother and the baby.

AI algorithms may also be used to for real time recognition of anatomical landmarks seen on US image. Those sophisticated AI algorithms may provide feedback signals such as name tags and transparent color overlays for each anatomical structure. US scan success rate of the entire image can also be evaluated with these algorithms (Fig 3, 4).

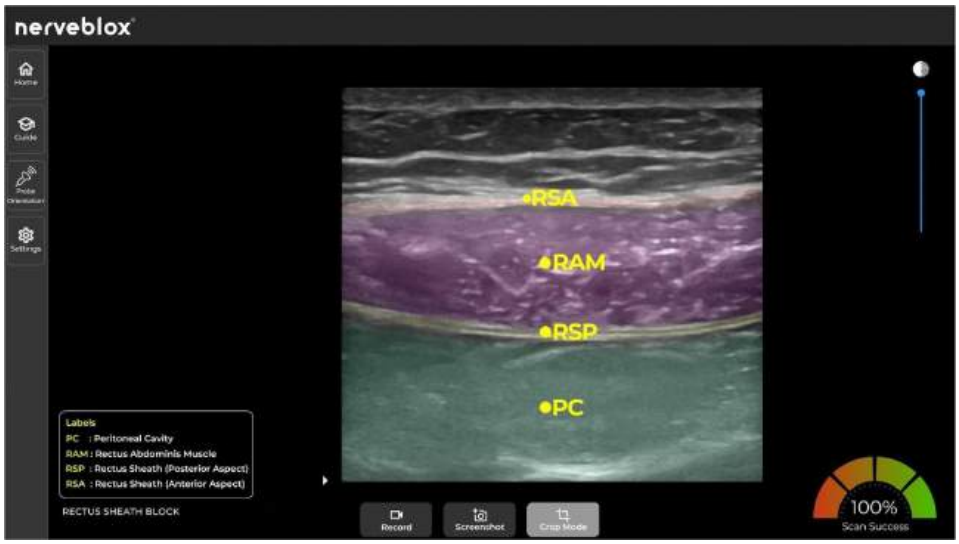


Figure 3. Display of 100 % scan success in US scanning is indicated in the right lower corner as a color scale. Related anatomical landmarks are indicated in the left lower corner as abbreviations (RSP: Rectus Sheath -Posterior Aspect, RSA: Rectus Sheath Anterior Aspect, RAM: Rectus Abdominis Muscle. PC: Peritoneal Cavity.). The images are obtained with a specialized AI software (Nerveblox, Smart Alfa Teknoloji San. ve Tic. A.Ş., Ankara, Turkey).

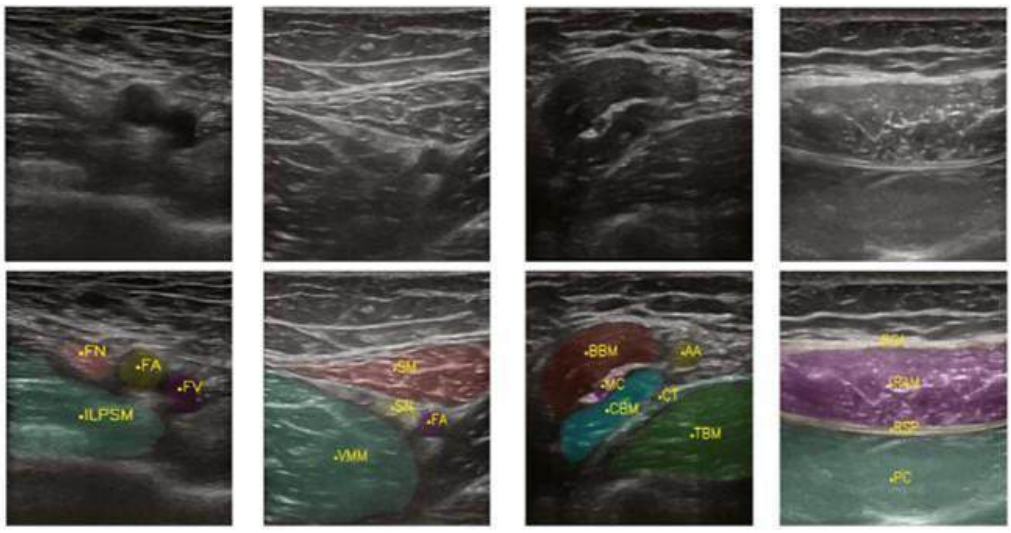


Figure 4. US images are demonstrated in pairs: Grey scale routine US images are displayed on the upper raw and AI processed images are displayed on the lower raw. Predefined related anatomical landmarks (abbreviated) are indicated within the AI processed images on the lower panel (Nerveblox, Smart Alfa Teknoloji San. ve Tic. A.Ş., Ankara, Turkey).

These applications have potential to improve performance of US scanning in training less experienced users and improving skills of experts. It could be especially useful for some interventional procedures such as US guided peripheral nerve blocks which used in different clinical settings for providing regional anesthesia in variable surgical procedures, and for pain control of acute injuries (12).

Additionally, the application of DL techniques to US image formation may provide high resolution images which are less blurry. To improve US image resolution hardware modifications can be made which are usually more complex and costly compared to software post-processing methods. Qualitative analysis of the reconstructed images generated using DL based software post-processing confirms that these methods significantly enhance image quality. These techniques provide high resolution images with better visual clarity to human visual system (5, 13).

Measurement and quantification of different organs and pathologies are important part of US imaging in monitoring disease conditions. Deep learning-based image processing algorithms used in quantification and computer-aided detection have now outperformed traditional engineering approaches. Another advantage of AI is that through computer assisted diagnosis AI may reduce physicians' workload and image interpretation time.

Computer aided ultrasound diagnosis

Computer Aided Diagnosis (CAD) has become one of the most popular research areas in radiology. It refers to the use of AI to assist radiologists in interpreting medical images. CAD can analyze images obtained through modalities such as X-ray, CT, MRI, and US. It can detect abnormalities, measure lesions, and assist in disease detection and classification. It provides a "second opinion" to support radiologists in image interpretation and reduces image reading time.

CAD mainly focuses on two aspects:

1. *Detection:* refers to the technology locating the lesion in an image, meaning an abnormality in the structure of a tissues or organ due to injury or disease especially the ones that are well circumscribed. It aims to simplify the visualization of abnormalities for medical staff.
2. *Diagnosis:* refers to identification of the potential diseases. It aims to provide additional support for the physicians to reach an accurate diagnosis.

The US based CAD systems operate in several steps including image preprocessing, image segmentation, feature extraction and selection, lesion classification as shown in Figure 5.

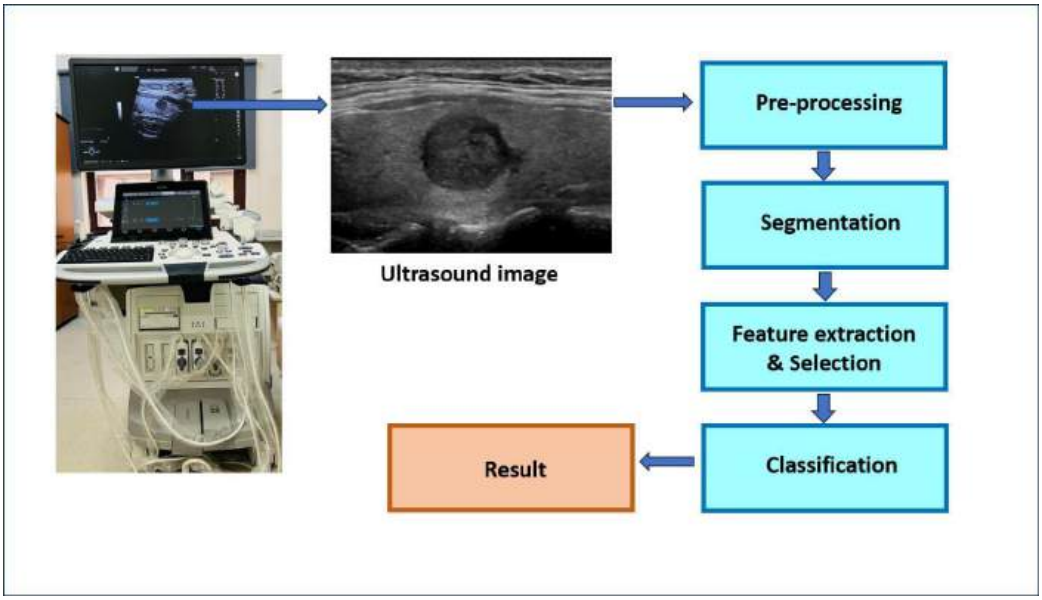


Figure 5. The general flowchart of CAD system for US

Numerous CAD systems have been developed in recent years especially for detection and classification of thyroid, breast, and liver diseases.

One of the most common research areas of AI in US is the thyroid. Thyroid nodules are extremely common in the adult population worldwide. US is the first line imaging tool for the diagnosis and management of thyroid nodules due to its well-known advantages such as its noninvasive nature. However, it has a limited ability to differentiate benign lesions from malignant ones. At present, the diagnosis of thyroid nodules mainly relies on biopsy, which is an invasive procedure. Even though thyroid biopsy is considerably safe, there is still risk of minor complications such as hemorrhage. It can also cause patient anxiety and increased cost. To reduce the operator dependency and improve diagnostic accuracy US-based CAD have been developed. These systems can automatically detect nodules from US images and help classify thyroid nodules according to their image characteristics (15). In a recent study examining 508 US images, including benign and malignant nodules, it was concluded that proposed method of DL could assist physicians` decision making, reduce the time for human participation, and improve the efficiency of diagnosis. Figure 6 shows the complex workflow for thyroid nodule classification. The output layer here defines the classification of thyroid nodule as benign or malignant (16).

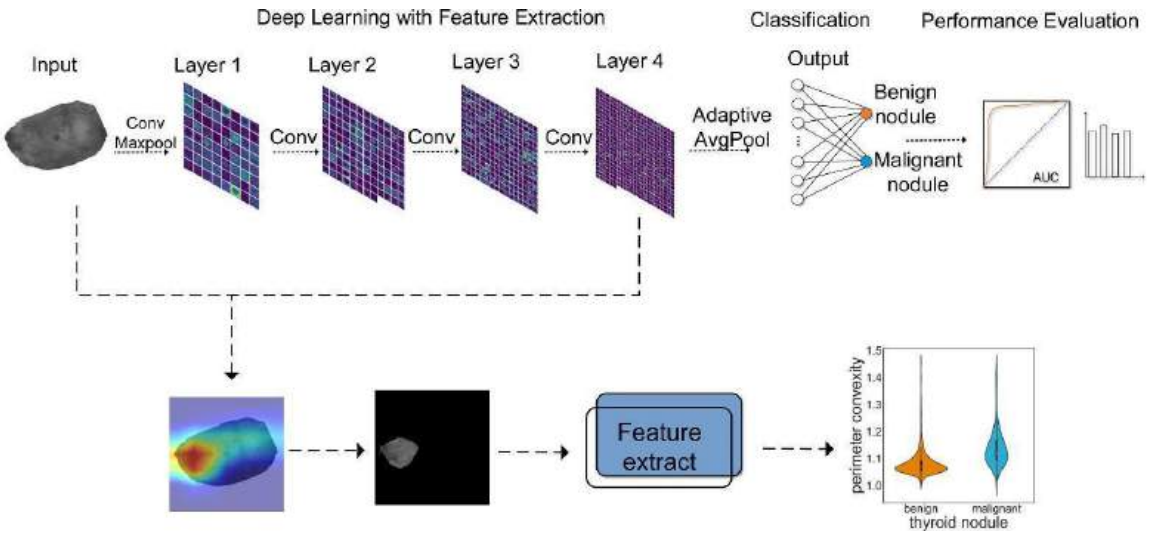


Figure 6: The complex workflow for thyroid nodule classification. After the convolution layers and pooling layers, the thyroid image features are extracted automatically. The output layer defines the classification of thyroid nodule as benign or malignant (Images are obtained from the study published in the Frontiers in Oncology by Yang J et al. (16).

Breast cancer is the most common cancer among women. CAD systems can be used to provide a second opinion to the radiologists in a cost-effective way. They can detect breast lesions and can be helpful in differentiating benign and malignant mass lesions. This way unnecessary biopsies can be avoided. CAD can also be used in early detection of malignant mass lesions by analyzing US features such as margin, shape and texture characteristics (4). In a recent study it has been concluded that CAD significantly improves diagnostic performance of radiologists with potential to reduce frequency of benign breast biopsies (17).

For the liver, US is the most commonly used and the first step imaging modality in the evaluation of diseases. Although liver biopsy is sensitive in diagnosis of liver diseases such as cirrhosis or mass lesions, it may cause complications such as bleeding, infection, and damage to some vital organs like lung or vessels. CAD supports detection and diagnosis of various liver lesions including benign and malignant ones, and helps monitor disease progression. It may also guide treatment planning (4).

In addition to these main applications of computer aided diagnosis in the US, several other AI applications have been investigated such as prostate, kidney, heart or fetus with some promising findings (14). Overall, CAD enhances diagnostic accuracy and supports decision making processes for different organs and diseases.

AI in ultrasound -unique challenges

In general, different factors contribute to the success of an AI application. The most important among them are large and well-defined data sets. The US has some challenges inherent to its technology, that make the integration of AI more complicated compared to other imaging modalities. The greatest challenges are operator dependency and image variability. US images

are obtained in real time, and the scanned area, image quality and probe type are selected by the user. US images may vary between different operators, and also between different US devices and different machine settings. In daily practice, images are not routinely archived in most of the clinics, only selected sample images are stored. This situation prevents the creation of large training data sets. Additionally, the interpretation of the US image can vary significantly between operators, resulting in potential errors in the "ground truth" data (4, 8). Because of these issues, progress in AI applications for the US is still behind that of AI-powered CT and MRI examinations. However, despite these limitations and challenges unique to US technology, studies published in recent years have shown promising results regarding the use of AI in US imaging.

Ethical issues On the other hand, clinicians remain ultimately responsible for the decisions they make regarding individual patients. Therefore, especially the experienced doctors are understandably hesitant about the integration of AI in their field. This situation is likely to change as the number of AI applications increases, and more clinicians begin to embrace technological advancements. This, in turn, will provide more feedback to the algorithm developers, feeding the improvement cycle and yielding more accurate results with DL.

Conclusion

AI technology has the potential to transform the field of US imaging, however unique challenges inherent to US technology such as user dependency prevents full automation in this modality. Therefore it is hard to imagine that AI will completely replace humans in US examinations anytime soon. However, the AI technology has potential to enhance US scanning performance, helping in the training of inexperienced users and improving the skills of experts. Also, AI-powered US systems may provide objective measurements and help in diagnosis of certain diseases. This approach will not only assist with the detection and accurate diagnosis of disease conditions, but also improve workflow of US scanning and help reduce healthcare costs. With ongoing advancements in AI and image processing algorithms, these techniques will be able to increase diagnostic accuracy of less-experienced users, ultimately enhancing the quality of patient care. In the near future, most probably every US machine will incorporate some level of AI automation and it will be routinely used in clinical practice.

Works Cited

MedlinePlus [Internet]. Bethesda (MD): National Library of Medicine (US); Available from https://medlineplus.gov/lab-tests/sonogram/

https://www.nibib.nih.gov/science-education/science-topics/ultrasound

Shung, K. Kirk. "Article Diagnostic Ultrasound: Past, Present, and Future." Journal of Medical and Biological Engineering 31, no. 6 (January 1, 2011): 371. https://doi.org/10.5405/jmbe.871.

Akkus, Zeynettin, Jason Cai, Arunnit Boonrod, Atefeh Zeinoddini, Alexander D. Weston, Kenneth A. Philbrick, and Bradley J. Erickson. "A Survey of Deep-Learning Applications in Ultrasound: Artificial Intelligence–Powered Ultrasound for Improving Clinical Workflow." Journal of the American College of Radiology 16, no. 9 (September 1, 2019): 1318–28. https://doi.org/10.1016/j.jacr.2019.06.004.

Tenajas, Rebeca, David Miraut, Carlos I. Illana, Rodrigo Alonso-Gonzalez, Fernando Arias-Valcayo, and Joaquin L. Herraiz. "Recent Advances in Artificial Intelligence-Assisted Ultrasound Scanning." Applied Sciences 13, no. 6 (March 14, 2023): 3693. https://doi.org/10.3390/app13063693.

Cambridge academic content dictionary :Cambridge University Press

Choy, Garry, Omid Khalilzadeh, Mark Michalski, Synho Do, Anthony E. Samir, Oleg S. Pianykh, J. Raymond Geis, Pari V. Pandharipande, James A. Brink, and Keith J. Dreyer. "Current Applications and Future Impact of Machine Learning in Radiology." Radiology 288, no. 2 (June 26, 2018): 318–28. https://doi.org/10.1148/radiol.2018171820.

Dicle, Oğuz. "Artificial Intelligence in Diagnostic Ultrasonography." Diagnostic and Interventional Radiology 0, no. 0 (January 11, 2023): 0. https://doi.org/10.4274/dir.2022.211260.

https://en.wikipedia.org/wiki/Deep_learning

Chi, Jianning, Ekta Walia, Paul Babyn, Jimmy Wang, Gary Groot, and Mark Eramian. "Thyroid Nodule Classification in Ultrasound Images by Fine-Tuning Deep Convolutional Neural Network." Journal of Digital Imaging 30, no. 4 (July 10, 2017): 477–86. https://doi.org/10.1007/s10278-017-9997-y.

Lin, Di, Athanasios V. Vasilakos, Yu Tang, and Yuanzhe Yao. "Neural Networks for Computer-aided Diagnosis in Medicine: A Review." Neurocomputing 216 (August 15, 2016): 700–708. https://doi.org/10.1016/j.neucom.2016.08.039.

Gungor, Irfan, Berrin Gunaydin, Beyza M. Buyukgebiz Yeşil, Selin Bagcaz, Miray Gozde Ozdemir, Gozde Inan, and Suna O Oktar. 2023. "Evaluation of the Effectiveness of Artificial Intelligence for Ultrasound Guided Peripheral Nerve and Plane Blocks in Recognizing Anatomical Structures." Annals of Anatomy - Anatomischer Anzeiger 250 (August): 152143–43. https://doi.org/10.1016/j.aanat.2023.152143.

Temiz, Hakan, and Hasan S. Bilge. "Super Resolution of B-Mode Ultrasound Images With Deep Learning." IEEE Access 8 (January 1, 2020): 78808–20. https://doi.org/10.1109/access.2020.2990344.

Shengfeng Liu, Yi Wang, Xin Yang, Baiying Lei, Li Liu, Shawn Xiang Li, Dong Ni, Tianfu Wang, Deep Learning in Medical Ultrasound Analysis: A Review, Engineering, Volume 5, Issue 2, 2019, Pages 261-275, ISSN 2095-8099, https://doi.org/10.1016/j.eng.2018.11.020.

Zheng, Yinghao, Lina Qin, Taorong Qiu, Aiyun Zhou, Pan Xu, and Zhixin Xue. 2021. "Automated Detection and Recognition of Thyroid Nodules in Ultrasound Images Using Improved Cascade Mask R-CNN." Multimedia Tools and Applications 81 (10): 13253–73. https://doi.org/10.1007/s11042-021-10939-4.

Yang, Jingya, Xiaoli Shi, Bing Wang, Wenjing Qiu, Geng Tian, Xudong Wang, Peizhen Wang, and Jiasheng Yang. 2022. "Ultrasound Image Classification of Thyroid Nodules Based on Deep Learning." Frontiers in Oncology 12 (July). https://doi.org/10.3389/fonc.2022.905955.

He, Ping, Wen Chen, Ming-Yu Bai, Jun Li, Qing-Qing Wang, Li-Hong Fan, Jian Zheng, et al. 2023. "Deep Learning–Based Computer-Aided Diagnosis for Breast Lesion Classification on Ultrasound: A Prospective Multicenter Study of Radiologists without Breast Ultrasound Expertise." American Journal of Roentgenology 221 (4): 450–59. https://doi.org/10.2214/ajr.23.29328.

Everything You Need to Know About Fake Handbags By Annie Tran

How do counterfeit handbag businesses impact the economy, consumer behavior, and brand strategies, and how can technology and policies address this issue?

Introduction

"It's just a fake handbag, no one will notice a difference", a saying that has been popular and prevalent for consumers, however these counterfeited bags are also known as "superfakes," these goods have become a significant issue which impacts negatively on the market of luxury items. Due to the imitation, these replicates are so precisely crafted that they are almost indistinguishable from the real item, this makes it an ideal product for consumers as it is affordable and a duplicate of the actual product. Because of this, it has gained a significant amount of attention and has increased in production in the market. Hence, the counterfeited product creates challenges and is a huge problem not just for brands but for the economies and consumers ethics of choices. The counterfeit handbag trade gains billions of dollars annually, but behind this success comes serious consequences, such as it takes away revenue from luxury brands, influences consumer behavior towards counterfeit goods without the knowledge of the issues that come with it, like unethical practices such as labor exploitation and environmental damage. Beyond this, counterfeit operations are often linked to organized crime, adding another layer of concern to this matter. In a world where fake can look just as good as real, does authenticity still matter? In this research paper, I will be focusing on different areas and stakeholders of the economy, like how counterfeited handbags can impact the economy, consumer behaviour, and how the brands companies find solutions to these issues keeping their brand's value of authenticity with the support of technology and policies set by the government.

Statistic and background

The counterfeit handbag industry has expanded into a multi billion dollar market, exerting a significant impact on the global economy and the luxury fashion sector. Counterfeit fashion items, including handbags, make up 51% of all counterfeit goods seized worldwide (Ennoventure, n.d.). In recent years, the global counterfeit market, including luxury goods, is now estimated to be worth $500 billion annually. This growth has been driven by advancements in manufacturing techniques for counterfeited goods, leading to the rise of "superfakes," an item that is almost nearly indistinguishable from the authentic luxury handbags. Skilled crafters are often employed to replicate details such as stitching, logos, and materials with precision. Additionally, the development of e-commerce and social media has made the distribution of counterfeit goods easier, making them widely accessible to consumers. Notably, a significant portion of buyers knowingly purchase these fake products, are drawn to the prestige of luxury branding at a significantly lower price.

Beyond market growth, the economic impact of counterfeit handbags is large. Counterfeiting costs the global economy approximately $320 billion annually in lost revenue for

legitimate luxury brands (Fashion United, 2018). As demand for these lower cost alternatives increases, the brands face declining sales, weakening their brand value. Moreover, the counterfeit industry raises significant ethical concerns. Many counterfeit goods are produced in sweatshops, a workplace where employees receive low wages and work in unsafe or exploitative conditions. Additionally, counterfeit manufacturing often disregards and ignores the environmental regulations, contributing to pollution, waste and environmental harm. Despite these consequences, a significant portion of consumers knowingly purchase counterfeit handbags, attracted by their affordability and the status associated with luxury branding. While some buyers remain unaware of the ethical and economic implications, the accessibility of these products continues to drive demand, further embedding counterfeit goods within the global market.

Manufacturing process

The production of counterfeit handbags is a sophisticated and highly organized process that copies legitimate luxury brand manufacturing while operating in an unregulated and unethical manner. Many counterfeit operations rely on migrant workers who work in unsafe and unregulated conditions (International Labour Organisation, 2024). These workshops often operate illegally, bypassing labor and safety regulations to minimize production costs. Despite these exploitative conditions, the counterfeit manufacturers have become increasingly skilled at replicating luxury handbags, employing advanced techniques to create near identical replicas.

The manufacturing process typically begins with the disassembly of genuine designer handbags. Counterfeiters reverse architect these products to replicate their structure, stitching, and material composition with precision. Depending on the quality tier, some counterfeit handbags are made from low-cost, low quality materials, while high end "superfakes" utilize premium leather, metals, and intricate detailing to closely mimic authentic luxury products. In many cases, counterfeit operations hire skilled artisans, sometimes former employees of luxury brands, they would apply their expertise to ensure that stitching, logos, and hardware are nearly indistinguishable from the originals.

The industry also operates within a global supply chain, sourcing materials and components from both legitimate and illegitimate suppliers. This allows manufacturers to blend authentic looking features with cost cutting alternatives, enhancing the realism of the final product. Once production is complete, counterfeit handbags are distributed through sophisticated networks, including street vendors and e-commerce platforms, effectively bypassing regulatory oversight and making these items easily accessible to consumers worldwide. With technological advancements further refining the replication process, counterfeit handbags continue to blur the lines between authentic luxury and replicates, posing ongoing challenges for luxury brands and policymakers.

The relationship between counterfeited goods and organized crimes

The counterfeit handbag trade is not just an economic and ethical issue but also a major contributor to organized crime and global security threats. Criminal corporations have

established a strong foothold in the counterfeit industry, using it as a profitable revenue stream to fund various illicit activities. The global value of counterfeiting is projected to reach approximately $3 trillion, with organized crime groups significantly contributing to this figure (ACVISS, 2024). The profitability and relatively low legal risks associated with counterfeit goods make them an attractive funding source for criminal organizations. Investigations have traced profits from counterfeit sales to groups engaged in drug trafficking, human trafficking, and even terrorism. Notably, funds from counterfeit goods sales have been linked to financing terrorist organizations such as Hezbollah and Al-Qaeda, with evidence suggesting that attacks like the 1993 World Trade Center bombing and the 2015 Charlie Hebdo attack were partially financed through counterfeiting operations (Vision of Humanity, 2024).

Beyond direct financing of illicit activities, the counterfeit handbag trade enables crime corporations to exploit vulnerable labor forces. Many counterfeit manufacturing operations function as today's sweatshops, where workers, often migrants, endure long hours with minimal pay in unsafe conditions (Independent, 2024). These exploitative practices help criminal organizations maintain low production costs while maximizing profits. Furthermore, counterfeit operations contribute to widespread tax evasion, depriving governments of substantial revenue that could otherwise be allocated to public services and economic development. This loss in government funds not only weakens economies but also pressures law enforcement's ability to combat counterfeit-related crimes effectively (ACVISS, 2024).

In addition to economic and labor exploitation, the counterfeit trade undermines legal systems by overwhelming intellectual property enforcement efforts. Law enforcement agencies face significant challenges in tracking and shutting down counterfeit operations, as these networks often operate across multiple countries with complex supply chains.The involvement of organized crime in the counterfeit industry has made regulation increasingly challenging, as resources are redirected toward combating counterfeiting instead of addressing other criminal activities. Ultimately, consumers who purchase counterfeit handbags may unknowingly contribute to this cycle of criminal funding and exploitation, reinforcing the need for stricter enforcement and greater awareness of the broader consequences of these items.

Luxury Brands Most Affected by Counterfeit Handbags

The counterfeit handbag industry impacts high-end luxury brands significantly, with some of the most prestigious fashion companies suffering significant financial and reputational damage. Brands such as Louis Vuitton, Gucci, Chanel, and Hermès are among the most affected by counterfeit handbag sales due to their strong brand recognition and exclusivity (Betsy Bags, 2024). The more recognizable and aspirational a brand is, the more likely it is to be targeted by counterfeiters who seek to take advantage of the high demand for luxury goods. These brands rely heavily on their perceived exclusivity and craftsmanship, making counterfeiting a direct threat to their market positioning.

The economic impact of counterfeiting on luxury brands is large, with billions of dollars lost annually due to unauthorized reproductions flooding the market. Counterfeit sales not only

divert revenue from legitimate brands but also diminish the value of authentic products. Consumers who unknowingly purchase fake handbags may become wary of buying luxury goods in the future, fearing that they might receive an imitation instead of a genuine item. This erosion of consumer trust can have long term consequences, damaging a brand's reputation and reducing demand for its products (Betsy Bags, 2024).

Geographically, companies based in the United States, France, Italy, Switzerland, and Germany are the most targeted by counterfeiters, with the U.S. leading as the largest source of affected brands (Statista, 2024). The luxury goods sector, particularly handbags, faces heightened vulnerability due to its high profit margins and global desirability. In addition to handbags, other high profile brands such as Rolex, Prada, and Michael Kors also experience significant counterfeit activity, highlighting the widespread impact of counterfeiting across the luxury industry. As counterfeiters continue to refine their techniques, luxury brands must invest heavily in anti-counterfeiting technologies and legal measures to protect their products and maintain consumer trust in their brand authenticity.

The impact of fakes on companies and how technology in combats the issue

The counterfeit handbag industry imposes severe financial and reputational damage on luxury brands, weakening their market presence and consumer trust. Counterfeiting costs high-end fashion companies billions of dollars in lost revenue annually, limiting their ability to invest in innovation, expansion, and new product development (Forbes, 2022). Beyond financial losses, counterfeit goods dilute brand equity by associating poor quality replicas with prestigious names such as Louis Vuitton, Chanel, and Hermès. Many consumers unknowingly purchase counterfeit handbags, mistaking them for genuine products, which creates confusion in the market and damages brand reputation. Additionally, luxury brands must allocate substantial resources to legal battles and enforcement efforts to protect their intellectual property, further straining their operations. Counterfeit infiltration into legitimate supply chains also complicates inventory management and distribution, while the availability of cheaper counterfeit goods intensifies competitive pressure, making it harder for luxury brands to maintain their pricing strategies (RFID Label, 2024).

To combat the growing threat of counterfeit handbags, luxury brands are leveraging advanced technology to enhance product authentication and security. One of the most effective solutions is blockchain technology, which provides transparency across the supply chain. By integrating blockchain, brands can ensure that each step of the production and distribution process is verifiable, allowing both businesses and consumers to authenticate products at every stage (Forbes, 2022). Additionally, artificial intelligence (AI) and machine learning are increasingly used to monitor online marketplaces, detecting and removing counterfeit listings before they reach potential buyers.

Another technological innovation being widely adopted is RFID (radio-frequency identification) technology, which embeds unique tracking tags into luxury handbags. These tags enable real-time authentication and help brands prevent counterfeit infiltration into legitimate

supply chains (RFID Label, 2024). Similarly, QR codes and NFC (Near Field Communication) tags allow consumers to verify the authenticity of their purchases using smartphones, providing a cost-effective and anti-counterfeiting measure. Luxury brands are also employing serialized holograms and security labels, which feature intricate, hard to replicate visual markers as an additional layer of protection against counterfeiting.

As counterfeiters continue to refine their techniques, luxury brands must stay ahead by integrating these technological advancements into their security strategies. By combining digital authentication methods with supply chain transparency, high-end fashion companies can strengthen consumer trust, reduce financial losses, and protect their brand integrity against the growing counterfeit market.

Psychology on why consumers buy counterfeit goods

The appeal of counterfeit luxury handbags goes far beyond their lower price tags. Psychological factors heavily influence consumer behavior, making counterfeits an attractive alternative for many shoppers. Luxury brands symbolize status, exclusivity, and success, yet for a large portion of the population, owning an authentic designer handbag remains financially unattainable. The desire to project wealth and prestige drives individuals toward counterfeit alternatives, allowing them to appear affluent without the hefty price tag. The association between luxury brands and social standing reinforces this behavior, as consumers see fake handbags as a shortcut to the status that high-end fashion conveys (Psychology But Make It Fashion).

Beyond the pursuit of status, the idea of belonging plays a critical role in the counterfeit market. In many social circles, possessing luxury goods signifies relevance and sophistication. Even when the item is fake, the mere appearance of wealth can be enough to foster social acceptance. The widespread popularity of "dupes" on social media has further normalized counterfeit purchases, particularly among younger generations who no longer see them as unethical. Instead, they view fake designers as an economical way to stay fashionable while keeping up with trends (Psychology But Make It Fashion). Seeing influencers openly uses replicas only strengthens this perception, making it increasingly difficult to distinguish between what is aspirational and what is simply imitation.

Consumers also justify buying counterfeits by weighing the benefits against the financial burden of luxury purchases. Many believe they can achieve the same aesthetic and social advantages without the expensive price, reinforcing the idea that counterfeits offer the best of both worlds. This internal reasoning allows buyers to suppress any guilt associated with supporting counterfeit markets. However, for those who do experience a moral dilemma, cognitive dissonance plays a role. Some resolve this conflict by convincing themselves that luxury brands overcharge for their products, making counterfeit purchases a form of financial save rather than ethical compromise (Fashion is Psychology).

The thrill of finding and purchasing a well made fake adds another layer to the psychological appeal. Scoring a high quality counterfeit that closely resembles an original

designer piece can provide a rush of excitement, similar to bargain hunting. This emotional high reinforces the behavior, making repeat purchases more likely. Social proof only amplifies this cycle, as consumers see peers and influencers embracing counterfeit culture without hesitation. As a result, the psychological barriers that once deterred consumers from buying fakes continue to erode, making counterfeit handbags not just an economic choice but a deeply ingrained psychological phenomenon.

Conclusion

The counterfeit handbag industry is more than just a problem for luxury brands. It affects economies, workers, and consumers alike. High-end fashion houses lose billions in revenue, while counterfeit operations thrive, often linked to organized crime and exploitative labor. At the same time, many consumers knowingly buy fake bags, drawn by the status they bring at a lower price. Social media and shifting attitudes have made counterfeits more acceptable, blurring the line between real and fake.

Luxury brands are fighting back with advanced technology, but the problem persists. While companies can take measures to protect their products, consumer choices shape the demand for counterfeits. Every purchase is a decision that either supports ethical production or fuels the counterfeit trade.

Works Cited

Andrew. “How Luxury Brands Protect Their Products Using RFID.” *RFID LABEL*, 23 Sept. 2024, www.rfidlabel.com/how-luxury-brands-protect-their-products-using-rfid

Francombe, Amy. “High-End Fashion Dupes Are Soaring Where Knock-Offs Never Could.” *Wired*, 13 Oct. 2024, www.wired.com/story/high-end-fashion-dupes-are-soaring-where-knock-offs-never-could/.

Frith, Maxine. “Introducing the Latest Accessory to Organised Crime: Fake Handbags.” *The Independent*, 29 July 2005, www.independent.co.uk/news/uk/crime/introducing-the-latest-accessory-to-organised-crime-fake-handbags-302491.html.

Goulet, Sénatrice Nathalie. “Counterfeiting: An ABC of Terrorist Funding.” *Vision of Humanity*, Orne Department (Normandy), 17 Oct. 2023, www.visionofhumanity.org/counterfeiting-an-abc-of-terrorist-financing/.

Jones, Rachel. “Council Post: The Impact of Fakes and How to Combat Them.” *Forbes*, Forbes Magazine, 12 Aug. 2024, www.forbes.com/councils/forbesbusinesscouncil/2022/02/08/the-impact-of-fakes-and-how-to-combat-them/.

Katz, Frances. “Do Counterfeit Bags Really Fund Terrorism?” *PurseBlog*, Frances Katz, 13 Dec. 2019, www.purseblog.com/report/do-counterfeit-bags-really-fund-terrorism/.

Multisiteadmin. “Genuine Handbag Sales Affected by Fake Handbag Sales.” *Betsy Bags*, 28 May 2020, betsybags.com/handbag-sales-fake/.

Pbmif. “The Psychology of Knockoff Fashion: What Drives the Desire for Luxury Goods?” *PBMIF*, PBMIF, 2 Nov. 2024, www.psychologybutmakeitfashion.com/post/to-yves-or-not-to-yves-the-psychology-behind-why-people-buy-knockoffs-and-all-that.

Ritcher, Felix. “Chart: U.S. Companies Most Affected by Counterfeiting | Statista.” *Statista*, www.statista.com/chart/17407/countries-most-affected-by-counterfeit-and-pirated-goods/. Accessed 10 Feb. 2025.

“Superfakes: Copycat Manufacturers Are Becoming Increasingly Skilled at Producing Knock-off Designer Handbags.” *ABC News*, Doc Louallen, Deborah Kim, Caroline Pahl, Lizann Robinson, Tara Guaimano, and Karen Ye, 29 Apr. 2024, abcnews.go.com/Business/superfakes-copycat-manufacturers-becoming-increasingly-skilled-producing-knock/story?id=109344382 .

Selvaratnam, Naomi, and Matt Henry. “It’s so Profitable, Even the Cop Stamping It out Admits This Is ‘the Crime I’d Get Into.’” *ABC News*, ABC News, 29 Feb. 2024, www.abc.net.au/news/2024-02-29/the-true-cost-of-fake-fashion-foreign-correspondent/103455512.

“The Dark Connection between Counterfeiting and Organized Crime.” *Acviss*, Acviss | Blog, 27 Mar. 2023,

blog.acviss.com/the-dark-connection-between-counterfeiting-and-organized-crime-clfqf5ei2455181jnzb82rvo78/.

West, Jane. “The Psychology of Knock-Offs.” *Fashion Is Psychology*, Fashion Psychology, 21 Sept. 2022, fashionispsychology.com/the-psychology-of-knock-offs/.

Zilber, Ariel. “Dior, Armani’s Pricey Handbags Made by Migrant Workers Who Make $2 an Hour, Cost $57 to Make: Prosecutors.” *New York Post*, New York Post, 4 July 2024, nypost.com/2024/07/04/business/dior-armanis-pricey-handbags-made-by-migrant-workers-who-make-2-an-hour-prosecutors.

Translated – Technology for Real-Time Analysis of Natural Sign Language with AI-Powered Translation for Empowering the Deaf Population

By Ayaan Jain

Abstract

Sign language is a tool for communication used by the deaf population. However, given the growing vocabulary size and variation between regions, there exists a communication barrier between the hearing and the deaf population hampering the quality of life and resources available to the deaf population. Researchers have attempted to break down this barrier using smart gloves, and AI-based sign language detection using real-time videos, and non-profit organizations have tried to popularize sign language by educating the hearing population on the same. In our following research, we tested various AI-based approaches for sign language detection that work effectively for sign language recognition using limited data points. The Cascaded Mediapipe-MLP model performed the best across datasets with an average of >90% accuracy. Additionally, we collected our own dataset and tested a first-person point of view (POV-1) detection to increase personalisation and reduce the number of downloads required. Through our research, we aim to empower deaf individuals with our tool and effectively communicate with the hearing population.

Introduction

Deafness and communication

Communication is one of the most basic ways that human beings can interact with each other, yet, at the same time, millions of people in different parts of the world have to confront serious barriers because of hearing disabilities. In a report presented by the World Health Organization, it has been estimated that more than 430 million people all over the world have hearing loss (1); out of these, 34 million are children (1). A large part of the deaf community depends on sign language as a main means of communication. Sign language is a tool for communication that uses hand gestures, facial expressions, and body movements. Fingerspelling is a part of sign language that involves spelling out words such as proper nouns letter by letter with specific handshapes. The problems faced by those with hearing loss require an in-depth study of the different types of deafness and their causes. Deafness may be from birth, that is, congenital, or it may be acquired later in life because of certain forms of illness, injury, or age factors. Consequences can be social, educational, and even professional, since one born deaf may experience life differently from a person who lost hearing later in life. In a study conducted by the National Deaf Center on US students, it was found that an estimated number of 237,000 students enrolled into college were deaf (2).

Literature review - Approaches to improve the communication gap between deaf and hearing populations

The issues of communication barriers remain intact for the deaf, especially during their interactions with hearing individuals who do not understand sign language. These obstacles can result in misunderstandings, social isolation, and limited access to important services and opportunities. Various efforts have been made to address these challenges. While technological advancements in hearing aids and cochlear implants have worked towards making life better for many (3), they are not accessible to everybody due to financial constraints (4). Awareness campaigns and policies by governments and organisations have also been launched to make sure full inclusivity and accessibility are considered. Despite all such moves, the gaps persist. Technological solutions usually lack natural, real-time communication, and awareness initiatives might not reach every section of society. In earlier studies, researchers have used a triboelectric smart glove for AI-enabled sign-language recognition (5) and leap motion controllers (6), however, a drawback to these solutions is that they can reach very high prices, which are not affordable by a vast majority of the Indian population. In order to provide an affordable solution, researchers have used a computer vision approach to classify the gestures signed (7). However, the generalisability of this approach to numerous gestures is yet to be tested. Additionally, deep networks require more data points to train efficiently and converge to a global minimum.

Our proposed robust solutions

Our proposed solution to reducing the communication gap involves developing an AI-powered sign language translator that is robust enough to work with any degree of background noise and a limited number of data points. By comparing the performance of various deep network models developed using Convolutional Neural Network (CNN) and the Object Detection frameworks (Cascaded Mediapipe-MLP), we scrutinised the algorithm that gave the best results for a limited number of data points and a greater number of classes. The models were trained and tested on datasets which not only include a second point of view, but a first-person point of view as well.

Need for a first-person point of view translator system

The major drawback for a second point of view is that a camera would need to be held up to face the signer which will not only be an inconvenience but also break the natural flow of the conversation. Therefore there is a need for a first point of view. Instead of requiring others to have the app and use it from a second point of view, only the signer would need access to it, ensuring consistent access to the software while minimising additional downloads. Our proposed solution is designed in such a way that it captures and translates sign language into words, both in speech and text formats, in real-time, enabling flawless communication between deaf and hearing individuals. The novelty of our proposed solution is that it is personalised, cost-efficient, requires minimal data points to train the model and can be extrapolated to any number of gestures.

Methodology

Overview

The main goal of this study was to develop an application that is capable of collecting personalised RGB datasets and then implementing different deep network models to analyse the data for the translation of finger spelling resulting in the classification of different alphabetical classes and finally the prediction of specific letters. The methodology consists of data collection, preprocessing, feature engineering, model architecture and selection, and finally evaluation metrics.

Data Collection

First-person point of view (POV-1)

As there were no pre-existing POV-1 datasets, we recognised the need for one and therefore created it ourselves. In order to collect data for our POV-1 dataset we utilised a camera setup that involved: a tripod attached to the signer at chest height; an iPhone 16 as our camera; and the gestures were performed at a distance of about 1 foot from the camera. Using the previously mentioned setup and our application for the collection, we collected about 200 images per class. As there are 26 alphabets we ended with 26 distinct classes and a dataset of 5,385 unique images. These images were produced using a simple, solid-colour background consisting of grey and white.

Second-person point of view (POV-2)

Although open-source datasets were already available, in order to maintain consistency with the POV-1 approach we felt it was necessary to create our dataset. The open-source dataset was obtained from Kaggle and was created by author Sowmya Iyer in 2020 (8). It contains 26 classes with a total of 3,331 images collected from 5 different signers including a variety of different backgrounds.

In order to collect the images for our POV-2 dataset we implemented a different method when compared to the POV-1 approach. This time instead of using a tripod and a phone camera, we used a webcam (Baytion 1080P USB Webcam) placed upon a stand at chest level, ensuring only the signer's hands were visible in the collected images. By changing the parameters and employing the same collection application, we produced another dataset including 26 distinct classes and 5,101 unique images.

To maintain cross-dataset consistency, we used the same gestures for the finger-spelling as the Kaggle dataset (8) for the collection of all the alphabets in POV-1 and POV-2.

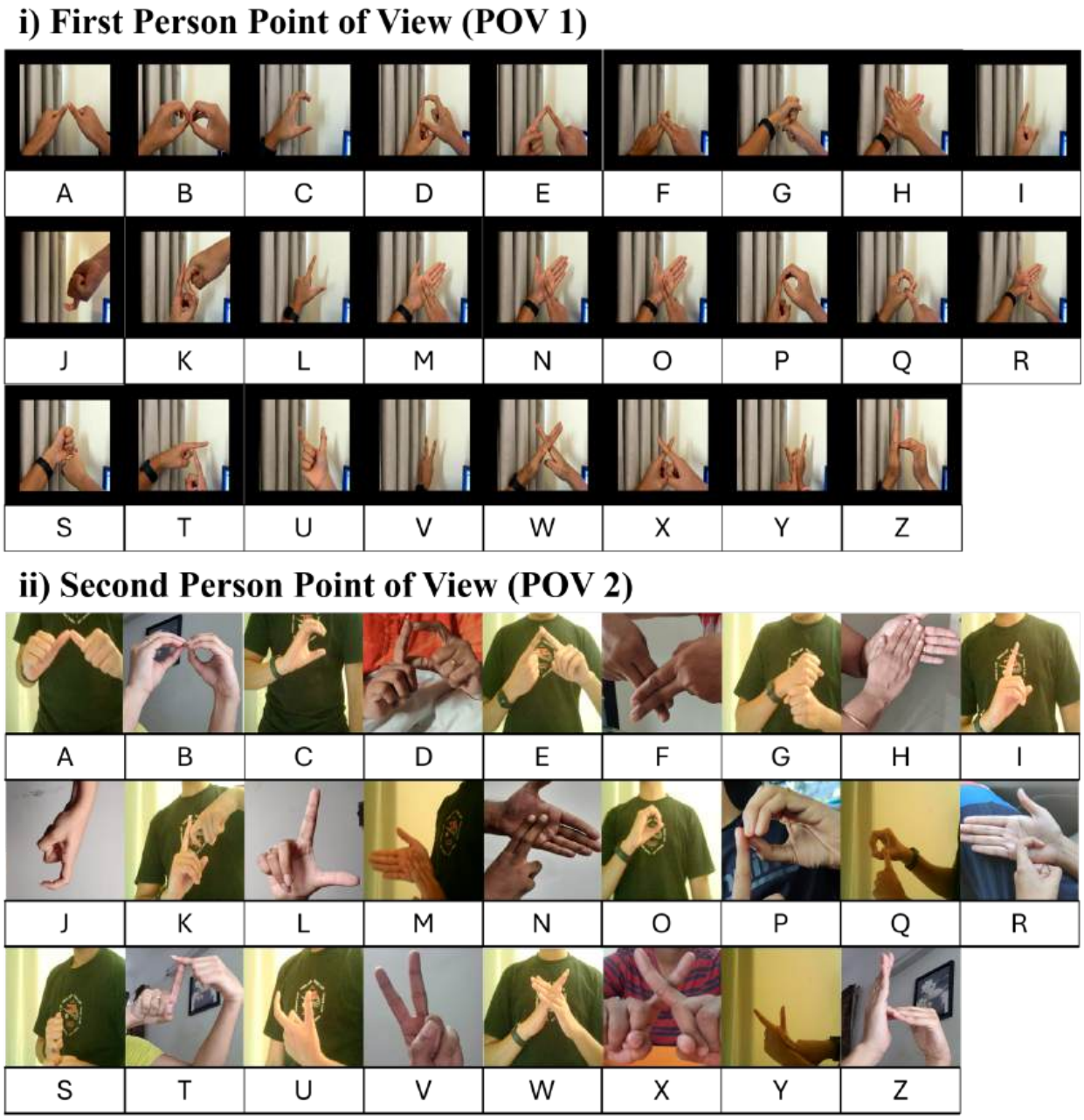


Fig. 1. POV-1 & POV-2 dataset i) includes POV-1 dataset made by signer ii) includes POV-2 dataset alternating between downloaded and own data

Data Preprocessing

Data cleaning involved manually removing blurry images through visual inspection. All the images were resized to 224 x 224 pixels, maintaining the aspect ratio using zero padding. For our CNN model, the images inputted were in the BGR format; however, for our Cascaded Mediapipe-MLP, the channels were permuted to RGB. The pixel values for each channel ranged from 0 to 255. The output from the model was a one-hot label encoded vector of 26 classes.

Model Architecture

In the study we conducted, we covered three distinct models. These include a SimpleCNN we developed, the pre-existing AlexNet architecture, and a Cascaded Mediapipe-MLP model we developed ourselves. We decided upon these models as we believed that they would yield the highest results in terms of accuracy. The SimpleCNN model is effective as it has a lower number of trainable parameters which means in real-time there is a lower latency in prediction. However, due to the model having a lower number of trainable parameters, there is a tradeoff between speed and accuracy. On the other hand, the AlexNet model has a higher number of trainable parameters resulting in greater accuracy but a slower speed in real-time prediction. The Cascaded Mediapipe-MLP model is the most efficient as it is pre-trained and only detects key points on the hand while simultaneously removing the background.

SimpleCNN

Our SimpleCNN architecture consisted of four convolutional layers and one fully-connected hidden layer. More details about the model are depicted in Fig 2. We performed a grid search to tune the hyper-parameters, testing our data with changing batch sizes [8, 32, 128] and learning rates [0.001, 0.01, 0.1]. After tuning we finalised using a batch size of 32 and a learning rate of 0.01 for POV-1, however, for POV-2 the hyper-parameters of the inbuilt “adam” optimiser were used. The weights of the model were randomly initialised and the “categorical cross-entropy” loss function was used. Using the tuned hyper-parameters, the model was run for 150 epochs, both for POV-1 and POV-2, with the stopping criteria being the number of epochs.

AlexNet

The AlexNet model (9) comprises 5 convolutional layers, 3 max-pooling layers after the first, second and fifth convolutional layers, 2 fully connected layers of size 4096 nodes, and one output SoftMax layer of size 26. For both datasets, POV-1 and POV-2, a batch size of 32 and an inbuilt “adam” optimiser were used. Similar to simple CNN, the weights of the model were randomly initialised, and the model was run for 150 epochs, both for POV-1 and POV-2, with the stopping criteria being the number of epochs.

Cascaded Mediapipe-MLP

The Medipipe Model is an object detection deep network that locates key points and features from RGB images (10). It takes an input of an RGB image and outputs the coordinates of the detected key points normalised with respect to the frame dimensions. The output of the hand key points from the mediapipe model was spatially normalised concerning the wrist coordinate and size of the detected hand. This processed array was fed as input to our MLP model with 1 hidden layer of 25 nodes and the output was redefined to a one-hot label encoded vector of 26 classes. More details about the model are depicted in Fig 2. To optimise the parameters, we performed a grid search. Therefore the model uses a batch size of 32, a learning

rate of “adaptive”, an activation value set to “tanh”, a hidden layer size of 50, an alpha value of 0.001, and finally an initial learning rate of 0.001.

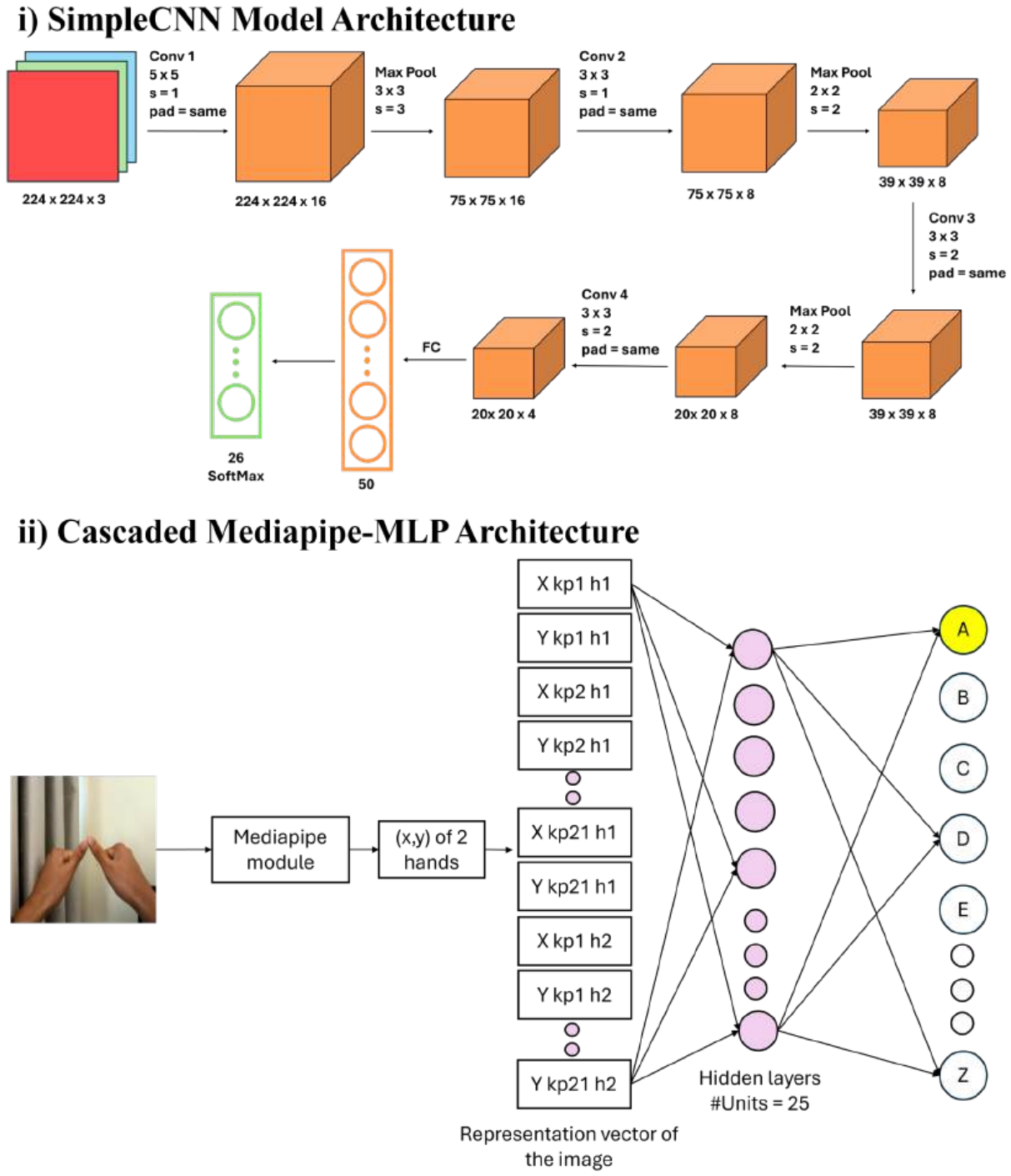


Fig. 2. SimpleCNN and Mediapipe-MLP Model Architectures

Training and Validation

For POV-1

As we only had one dataset for POV-1, we trained and validated the model on the same dataset. For the SimpleCNN model and AlexNet, we used an 80/20 validation split which means the model was trained on 80% of the dataset and validated on the remaining 20%. However, the approach used for the Cascaded Mediapipe-MLP model was different. We implemented a Stratified K-Fold cross-validation, which works slightly differently. The data set is divided into K, which is equal-sized folds, in our case 5. Each fold contains roughly the same percentage of samples for each class, preserving the class distribution of the dataset. It then trains the model on 4 of the folds; the remaining one is used as the test set. This is repeated for the number of folds, each time changing which set is the test set. As previously mentioned we only had one dataset for POV-1, therefore we were not able to test the model on an unseen dataset, but only train and validate it.

For POV-2

The method we used for training and validating the model for POV-2 was that of POV-1. For the SimpleCNN and AlexNet models, we used the 80/20 validation split and for the Cascaded Mediapipe-MLP, we used the Stratified K-Fold with 5 equal-sized sections. However, as we had two datasets for POV-2, we were able to test the model. We trained all the models on the downloaded POV-2 dataset and then tested it on the POV-2 dataset that we created. To maintain consistency with the POV-1 approach, we also trained and validated all the models using the created POV2 dataset following the same approach as POV-1.

Evaluation Metrics

Model Selection Metric

The models were compared and selected based on two metrics, the validation accuracy of the model and the loss value. The accuracy of a model can be defined by:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}$$

A categorical cross-entropy loss function was used to train and evaluate the model. It can be defined by:

$$L(y, \widehat{y}) = -\sum_{i=1}^{C} y_i \log(\widehat{y_i})$$

Where,

- $L(y, \widehat{y})$ is the categorical cross-entropy loss
- y_i is the actual class from the one-hot encoded vector

- $\widehat{y_i}$ is the predicted probability of the class
- C is the total number of classes

Real-Time Performance Metric

Latency is the number of frames processed in real-time per second. An average was calculated after running the model for one minute to calculate the latency. The best-performing model based on accuracy was deployed in real-time and the average latency was calculated after running the application for one minute. The experiments were run on a computer with an AMD Ryzen 5 5600X 6-Core Processor running at 4.2GHz using 16GB of RAM on Windows 10.

Detection efficiency of Mediapipe model

To calculate the detection efficiency of the model, the formula used is defined by:

$$Detection\ Efficiency\ = \frac{1}{C}\sum_{i=1}^{C}\frac{N_i^H}{N_i^{total}}$$

Where,

- C is the number of classes
- N^H is the number of images with H hands being detected
- N^{total} is the total number of images for class i

After calculation, the Mediapipe model's detection efficiency was 93% for POV-1 and 94% for POV-2. To reduce false negatives, the model's Minimum Detection Confidence and Minimum Tracking Confidence were both set to 0.05.

Results

Overview

The results section consists of Data Preprocessing, Model Performance Comparisons, and Real-Time Deployment.

Data Preprocessing

Through manual visual inspection, blurry and unclear images from each class were removed in order to increase the prediction accuracy of the model. An estimate of about 15 to 20 images were removed from the beginning of each class. As mentioned in the methodology section, the detection efficiency of the Mediapipe model was 93% for POV-1 and 94% for POV-2. However, this was an average; the detection efficiency for certain classes was lower than 30%, including M, N, R, and Y, as the model was not able to identify the hands as they were overlapping.

Model Comparisons

The three models being compared were AlexNet, our SimpleCNN, and the Cascaded Mediapipe-MLP model. The models were trained and tested against three different datasets: our POV-1 dataset, the downloaded POV-2 dataset, and the POV-1 dataset.

	Models	Training Accuracy	Validation Accuracy	Testing Accuracy
POV-1	AlexNet	98.89	41.69	-
	SimpleCNN	98.46	73.44	-
	Cascaded Mediapipe-MLP model	100	**99.40**	-
POV-2 (Own Dataset)	AlexNet	99.20	75.32	-
	SimpleCNN	99.47	**99.71**	-
	Cascaded Mediapipe-MLP model	99.60	99.59	-
POV-2 (Downloaded & Own Dataset)	AlexNet	53.16	35.14	7.88
	SimpleCNN	75.60	53.20	10.86
	Cascaded Mediapipe-MLP model	96.30	**84.14**	**77.58**

Table 1. Results of the Model Comparison between AlexNet, SimpleCNN, and Cascaded Mediapipe-MLP model

From Table 1 we observed that the Cascaded Mediapipe-MLP model works robustly across all the datasets. Furthermore, it had the highest testing accuracy when trained on the downloaded POV-2 dataset and tested on our POV-2 dataset. This proves that the Cascaded Mediapipe-MLP model works well at generalising across different datasets.

Real-Time Deployment

As the Cascaded Mediapipe-MLP was returning the best results, it was employed in an application using the Tkinter library to create a real-time detection application. The application features a way to capture new datasets. Using the application 26 classes for both POV-1 and POV-2 were captured, however, more classes can be included based on the user's requirement. The model can be trained again on the updated dataset and then deployed for prediction. It is to be noted that the time the model takes to be trained depends on the number of input classes and images. The model took approximately an hour to train 26 classes, each including about 200

images. The application features a set space which outputs the predictions based on the saved folder names. Real-time deployment does include lag, and the latency was calculated to be 12 frames per second.

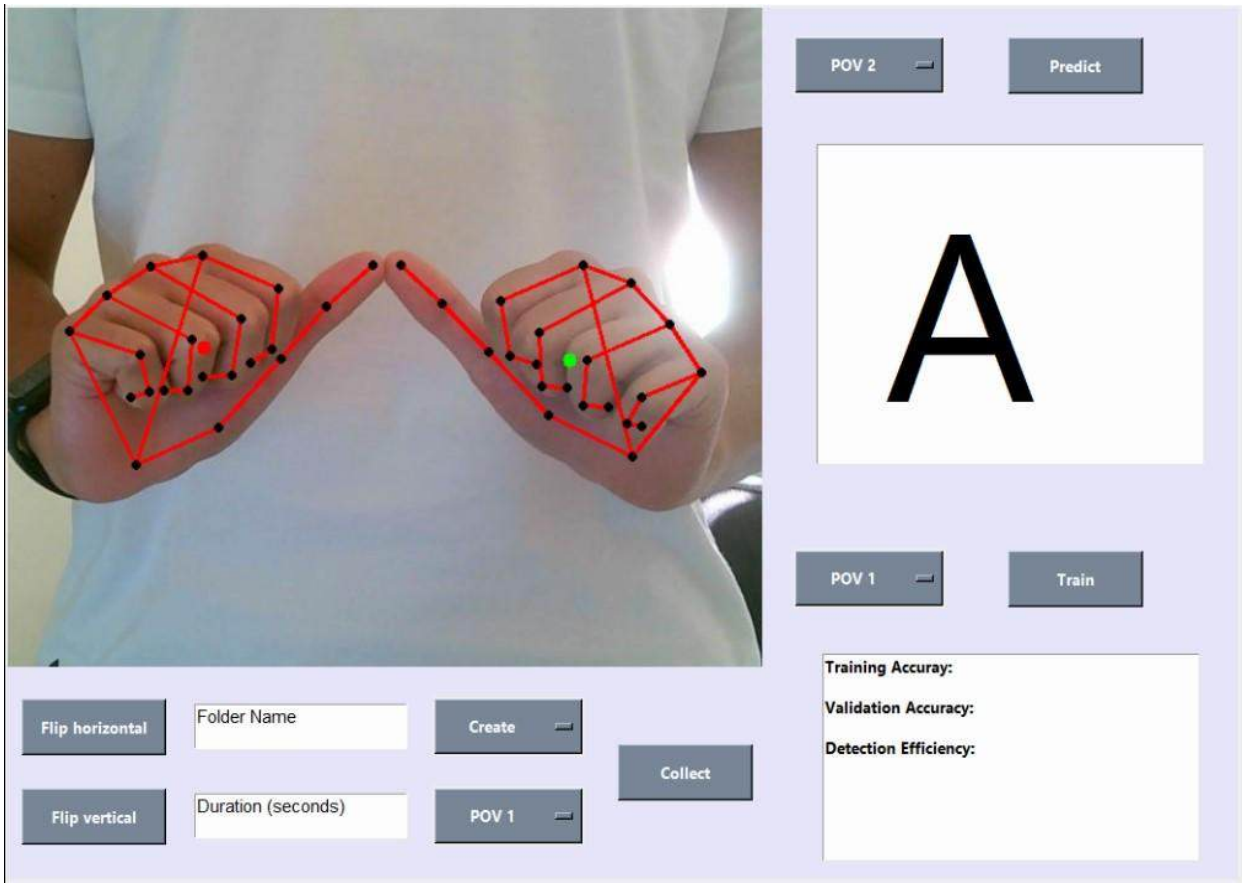


Fig. 3. Picture of the application to present prediction

The flip buttons visible in Fig.3. allow the user to change camera orientation to be best suited for their needs. Furthermore, there is an option to either create a new class or add to an existing one. Lastly, there is a dropdown menu which allows users to collect, train, or predict on either POV-1 or POV-2. A demonstration of the application is present in the following link (11).

Discussion

Summary

Three different methods were used to test our hypothesis: which approach works best for a first-person point of view? We found that the Cascaded Mediapipe-MLP solved our problem with the greatest level of accuracy. The reason why the Cascaded Mediapipe-MLP operates more efficiently compared to the SimpleCNN or AlexNet models is due to the two-step approach the model takes. Firstly, it performs image segmentation to detect the hand's key points. Then, using the identified key points, the MLP is trained separately to predict the classes. In contrast, the

CNN models consider the whole image, including the background noise, resulting in many distracting features being included within the training data. Furthermore, CNNs are better suited for larger datasets: if there were more training datasets available, the CNN models could learn the relevant features more effectively.

Limitation

Although the Cascaded Mediapipe-MLP returns quite a high prediction accuracy, there is a drawback to this model. If the hands overlap, the detection efficiency of the Mediapipe model drops drastically. For example, in our dataset, we observed this scenario in our dataset for the alphabet Y. We believe this is due to the Mediapipe model requiring both wrists to be visible to have confident predictions. This issue can be resolved by slightly modifying the gestures to ensure both hands are clearly visible.

Implementation Method

Nowadays, smart glasses are equipped with a display, camera and in some cases even speakers. A method of implementation for our proposed solution, a sign language translator, is to integrate the algorithm of the translator into the software of smart glasses. If a member of the deaf population would like to communicate with the hearing population, this can be achieved through the POV-1 system. The signer would be wearing the glasses and signing as normal, while the camera on the glasses picks up the signs and outputs the recognised gesture's translation through the speaker. On the contrary, if a member of the hearing population is wearing the glasses, it would carry out the translation using the POV-2 system. It works in the same manner as that of POV-1.

Future Directions

In order to further improve the algorithm, the number of classes inside the datasets could be increased. Currently, the dataset only has 26 classes which only includes the letters of the alphabet. Increasing the number of classes would allow full words to be added and would convergently validate the dataset. It would also allow full sentences to be translated, removing the cap of just fingerspelling. Additionally, to bridge the gap between hearing and deaf individuals even more, the model could be developed to include a method of translating speech or written text to sign language.

To conclude, *TRANSLATED* is one step closer to removing the divide between deaf and hearing individuals, establishing seamless communication cost-effectively and naturally.

Works Cited

World Health Organisation, “Deafness and hearing loss”, February 2024

Bloom, C. L., & Palmer, J. L. (2023). Undergraduate enrollment of deaf students in the United States (2019–2020). National Deaf Center on Postsecondary Outcomes, The University of Texas at Austin.

Cohen, S. M., Labadie, R. F., Dietrich, M. S., & Haynes, D. S. (2004). Quality of Life in Hearing-Impaired Adults: The Role of Cochlear Implants and Hearing Aids. Otolaryngology-Head and Neck Surgery, 131(4), 413–422. doi:10.1016/j.otohns.2004.03.026

Stephanie W, How Much Do Cochlear Implants Cost In 2024?, Forbes Health, March 2024

Wen, F., Zhang, Z., He, T. *et al.* AI enabled sign language recognition and VR space bidirectional communication using triboelectric smart glove. *Nat Commun* 12, 5378 (2021). https://doi.org/10.1038/s41467-021-25637-w

D. Naglot and M. Kulkarni, "Real time sign language recognition using the leap motion controller," 2016 International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2016, pp. 1-5, doi: 10.1109/INVENTIVE.2016.7830097.

Zhang, J., Bu, X., Wang, Y. et al. Sign language recognition based on dual-path background erasure convolutional neural network. Sci Rep 14, 11360 (2024). https://doi.org/10.1038/s41598-024-62008-z

https://www.kaggle.com/datasets/sowmyaiyer/isl-image-classification-with-a-background-clutter

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. Commun. ACM 60, 6 (June 2017), 84–90. https://doi.org/10.1145/3065386

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, Matthias Grundmann, MediaPipe: A Framework for Building Perception Pipelines, June 2019, https://doi.org/10.48550/arXiv.1906.08172

Link to the demonstration video: https://www.youtube.com/watch?v=6VXKH48hCmk