

|CURIEUX|

ACADEMIC JOURNAL

February Issue

Part 3 Issue 59

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

Page 6: Unraveling Anxiety: Exploring the Biological, Psychological and Social Roots of Anxiety Disorders By Arabela Vicelich

Page 19: Improvement of GraphCast Temperature Field and Prediction of Mid-Latitude Fronts Based on PINN Hard Constraints By Taoyu Pan

Page 36: Understanding the Parents and Kids By Noah Ikenberry

Page 41: Biomimetic Conductive Scaffolds Integrating Metal-Organic Frameworks and Piezoelectric Polymers for Wound Healing by Epithelial Electric Field Modulation By Noah Offiah

Page 52: Integrating microRNA Therapeutics and Nanotechnology for Precision Treatment of Human Diseases By Sharon Chang¹

Page 69: Reliable Gaze-Driven Medical Communication Boards Using Consistency Loss with Conjugate Gaze By JoonWoo Park and Hyunggyu Kim

Page 78: Rethinking Spatial Design for Elderly Communities Across Contexts: The Case of Singapore and Montgomery County, Maryland, United States of America By Aaron T. Chuang

Page 94: Predicting Sentiment Score in Social Media Trends Using Artificial Intelligence By Rishma Desai

Page 105: Ukraine's Economic Recovery: Sectoral Pathways to Post-War Growth By Alex Shymkiv

Page 120: To What Extent Can AI-Driven Predictive Tools Improve Agricultural Outcomes for Farmers in Indian States Vulnerable to Unpredictable Conditions? By Nikunj Garg

Page 124: Legal Loopholes in Theory and Practice: Historical Origins, Ethical Implications, and the Problems between Justice, Interpretation, and Exploitation By Elizabeth Wong

Page 140: The Fascinating Use of AI in Marketing By Ishaan Abhyankar

Page 151: Do Behavioral Biases Hit Harder in Crypto? A Comparative Literature Review of Behavioral Biases in Crypto and Stock Markets By Aydin Alp Erdev

Page 166: The Future of Real Estate: Tokenization and Fractional Ownership By Matheo Zuluaga

Page 174: Cyanobacteria: A Potential Sustainable and Efficient Source of Biofuel and Biodegradable Plastic By Pham Minh Tri

Page 187: Liberalism for the Colonised: French Occupation and Ideology in the Case of Algeria By Emma Shen

Page 192: The Influence of Socioeconomic Status on Legal Outcomes: The Interplay Between Economic Capacity and Access to Justice By Luna Han

Page 212: Relationship Between Parental Education and Financial Literacy of Youth in India By Advay Gupta

Page 228: Exo Ranker — A Fully Client-Side Visual-Analytics Framework for Earth-Similarity Assessment of Confirmed Exoplanets By Edward Xiong

Page 236: Learning Complementary Biomarkers: 2.5D ViT and Cross-Attention Fusion of Amyloid-PET and Tau-PET for CDR-level Alzheimer's Classification By Yicheng Lu, Michelle Xie

Page 245: To What Extent Does the Imposition of Value Added Tax Decrease Inequality? By Amelie Wu

Page 261: Implications of Genetic, Environmental, and Immunological Factors in Oral Lichens Planus and Their Insight on Therapeutic Interventions By Anshika Alamuthu

Page 273: Biological and Clinical Effects of Genetic Mutations in Female Associated Cancers By Mira Rakesh

Page 287: Holy Causes and Hidden Motives: Power and Faith in the Thirty Years' War Introduction By Miles Snyder

Page 298: Analyzing the Effects of Amman's Government Intervention On Fuel Consumption By Omar Khoury

Page 307: The Future of Human Gene Editing Through CRISPR/Cas9 By Samita Golla

Page 320: From Tool to Crutch: The Thin Line in AI-Enhanced English Learning By Siqu Fan

Page 331: Comparative Study of Parametric and Non-Parametric Models in Crop Recommendation By Krishna Ramakrishnan

Page 338: A Comprehensive Review of Depression: Risk Factors, Anatomy, and Treatment By Tanvi Shrowty

Page 352: Rewiring Hope: The Promise of Adult Neurogenesis-Based Regenerative Therapies By Avanish Srinivasan

Page 375: How Can Marketing Interventions Be Designed to Increase Attendance at High-School Athletic Events? By Roman Davis, Shemal Doshi

Page 385: Wick Rotation and Applications By Atharv Srivastava

Page 395: Predicting Type 2 Diabetes Using Machine Learning: Insights from Feature Importance and Clinical Symptom Analysis By Luke Choi

Page 403: Microloans, Fintech, and Female Entrepreneurship: A Conceptual Study of Women-led MSMEs in Kazakhstan By Maren Webb

Page 413: The Effects of Neurologic Music Therapy Techniques on Patients with Cerebral Palsy By Zachary W. Ong

Page 424: To What Extent Did Ideology and Propaganda Contribute to Mao's Rise to Power? By Xinyou Chen

Page 434: How Does the Brain's Perception of Episodic Memory Change over Time? By Arjun Balasundaram

Page 445: Using AI to Predict the Risk of Cervical Cancer By Bridget Parent

Page 458: Electromagnetic Interference (EMI) from Consumer Devices and Its Impact on Implantable Pacemakers with AI for Signal Classification By Gia Bhati

Page 479: A Vision-Based Framework for Automated Biomechanical Analysis and Feedback in Cricket Fast Bowling By Subinay Kalya

Page 501: Losing Quechua and Oneself: The Mental Health Impacts of the Extinction of Languages By Marco Kang Pastrana

Page 509: Why Does Stirring Hot Tea Make It Cool Faster? — A Geometric Study of Surface Area By Jimin Lee

Page 517: Artificial Intelligence in Alzheimer's Disease GWAS By Joy Lukama

Page 531: High School Educator Perceptions on Implementing Italian Education Practices Into the American Traditional Education System By Paige Bruskrud

Page 549: The Effects of Music Genres on EEG Waves By Isabelle Jiang

Page 556: Costs and Care for AD/ADRD Patients: Quantitative Analysis of Insurance Premium and Quality of AD/ADRD Treatment (2017-2021) By Kaavya Veligeti

Page 570: The Impact of Childhood Trauma on the Development of Gambling Behaviors in Young Adults By Imad Chaudhary

Page 575: Cost-Benefit Analysis of Climate Change Adaptation Strategies in Portugal, Mexico, and South Korea By Ramilya Makhmudova

Page 590: Analyzing Stock Market Behavior: Empirical Evidence From the United States By Zachary Leeds

Unraveling Anxiety: Exploring the Biological, Psychological and Social Roots of Anxiety Disorders By Arabela Vicelich

Abstract

Why do some people seem wired for anxiety while others remain unaffected? The answer lies in a complex web of biology and life experiences. This literature review examines the most significant biological, psychological and social factors contributing to the development of anxiety disorders. In particular, it discusses how specific biological factors such as genetic predispositions, neurotransmitter imbalances, and altered brain structure make individuals susceptible to anxiety based on existing research. At the same time, there are psychological influences – such as early life trauma, chronic stress, personality traits, and distorted cognitions or thinking patterns – which frequently determine how and when these symptoms surface. Social factors such as family dynamics, societal economic status, cultural pressures, and support availability further add to both risk and protective factors. Seeking to synthesize studies in neuroscience, psychology and psychiatry, this literature review focuses on the links between nature, nurture and environment with respect to anxiety. It contends that biological risk factors provide the foundation, but psychological and social experiences often determine the long-time severity and persistence of anxiety disorders. By understanding this dynamic relationship, this review hopes to contribute towards a more equitable biopsychosocial view of mental health. This perspective can help improve early identification and influence more personalized, effective approaches to treatment. Finally, the resolution of inherited, experiential, and social factors from which the anxiety emerges is also necessary to piece these all together into a complete picture as to how anxiety has developed for a person and how it can be managed.

Keywords Anxiety, biological factors, psychological factors, social factors, mental health, bio-psycho-social model

Introduction

For many people experiencing anxiety, it's not just occasional nerves – it's a persistent struggle that impacts relationships, work, and overall quality of life (Kessler et al.). Anxiety disorders are among the most common of mental health problems worldwide, with approximately 300 million sufferers in total and yet, their root causes remain complex and poorly understood (Domschke and Maron). Anxiety is a psychological disorder that involves excessive fear, apprehension, and worry, and can interfere with daily functioning and overall wellbeing (Kalin). To make sense of this complexity, this review applies the bio-psycho-social model, which views anxiety as the result of interacting biological, psychological, and social influences (Buckner et al.). This model suggests that no single factor fully explains mental health conditions, but rather, biological aspects such as genetics and brain chemistry, psychological elements like thought patterns and coping mechanisms, and social factors including relationships and environmental stressors all work together to shape a person's experience of anxiety.

Biologically, studies have indicated that anxiety disorders often run in families,

suggesting a strong genetic component (Hettema et al.). Studies have identified genes related to neurotransmitter systems, such as serotonin and dopamine regulation, which may contribute to heightened vulnerability by influencing how effectively these chemicals transmit signals in the brain (McKay et al.). For instance, imbalances in serotonin levels can affect mood stability and stress responses, increasing the likelihood of developing anxiety (McKay et al.). Additionally, neuroimaging has revealed that overactivity in the amygdala – a brain region central to processing fear and threat – is a consistent finding among individuals with anxiety disorders (Davis). Imbalances in the hypothalamic-pituitary-adrenal (HPA) axis – which controls the stress response – further demonstrate how biological systems can predispose individuals to chronic anxiety (Purves et al.).

Psychologically, anxiety is strongly associated with cognitive processes. Anxious individuals tend to use more of these maladaptive thinking styles such as catastrophizing and overestimation towards threat (Narmandakh et al.) Early experiences, such as childhood adversity or trauma, have also been associated with long-term difficulties regulating fear and stress, resulting in an increased risk for anxiety in later life (Malan-Muller et al.). In addition, psychological models such as classical conditioning explain how neutral situations become fear triggers that maintain a cycle of avoidance and anxiety (Davis).

On the social level, environmental stressors have a significant impact on maintenance of anxiety. Adolescent peer rejection, bullying and social isolation have been demonstrated to increase the likelihood of developing social anxiety disorders later in life (Barnao & Ward). Stressful life events in adulthood, including loss of employment, financial tightness and relationship problems are also important risk factors for the onset of anxiety and relapses (Ryu and Fan). Nevertheless, there are cultural and social pressures such as stigmatization of mental issues and excessive demands to be successful that can facilitate the symptoms by limiting access to support systems and fostering strong perceptions of incompetence (Łoś and Waszkiewicz).

An example of a study that applies the bio-psycho-social model on social anxiety and substance use is by Buckner et al. Biological mechanisms included physiological arousal; psychological factors involve fear of evaluation, negative mood, avoidance, and overthinking after social events; and social influences included peer norms, parental approval, the situations in which substances are used, such as drinking alone or before stressful social events. This study shows socially anxious individuals often use substances to manage subjective anxiety, even when physiological effects are minimal. By examining these interacting factors together, the review demonstrates the value of the bio-psycho-social model for understanding why anxiety increases vulnerability to substance misuse and informs more effective interventions.

Overall, these findings emphasise that anxiety arises from the interaction of multiple factors, rather than from a single cause. Instead, the bio-psycho-social model offers a more complete conceptualization that considers that these genetic vulnerabilities interact with learned cognitive patterns and environmental stressors to influence how anxiety arises and is sustained. This review will therefore evaluate the most significant biological and psychological

factors influencing adult anxiety, while also acknowledging the broader social context that amplifies their effects.

To organize this literature review, it will begin with an overview of the biological underpinnings of anxiety by highlighting genetic predispositions, brain chemistry, differential brain function, and physical health factors that lead to heightened vulnerability. It will then consider the psychological factors, including maladaptive thinking patterns, early life experiences, and learned behaviors that shape how anxiety develops and persists. Thereafter, social vulnerabilities such as family dynamics, cultural expectations, peer relationships, and societal pressures that may buffer or worsen anxiety symptoms. At the end of this literature review, these threads will be woven together in a conclusion that points out how this integration of biological, psychological and social factors offers a more comprehensive approach to understanding anxiety disorders, as well as treatment directions with potential for being more effective or individualized.

Biological Factors Contributing to Anxiety Disorders

Biological factors such as genetic vulnerabilities, brain chemistry, differential brain activity, neurophysiological imbalances, and physical health play a critical role in increasing an individual's vulnerability to developing and increasing the symptoms of anxiety disorders. There is strong evidence that anxiety runs in families, largely because of genetics. Hettema et al. note that among twin studies, it is common to find that diverse types of anxiety disorders share the same genetic risk factors and suggest that this finding may explain why individuals sometimes develop more than one disorder simultaneously. Crucially, carrying the genetic risk does not ensure that a person will develop anxiety: one twin could develop an anxiety disorder while the other does not, based on differences in environmental exposures, stress and life experiences. This occurs because the same genes that drive something like high sensitivity to stress or a propensity for becoming overly reactive in response to fear cues can be manifested in a variety of ways, leading to multiple styles of anxiety in the same person. Hettema et al. also point out that two of the largest genetic factors appear to be involved in multiple subtypes of anxiety disorders: a factor associated with panic disorder, generalized anxiety disorder, and agoraphobia; and a separate one more closely linked to phobias, such as fear of animals or particular situations. Social anxiety is in the middle because it seems to be affected by both sets of genes – some that are associated with panic disorder, generalized anxiety and agoraphobia, some linked to phobias (Hettema et al.) This combination goes a long way to explaining why social anxiety can overlap with other forms of anxiety.

Anxiety can also appear in families even if children are not raised by their biological parents, demonstrating that genetic vulnerability is a key factor aside from the environment (Davis). Genes alone do not cause anxiety, they merely predispose one to anxiety, meaning individuals who have these genes are more susceptible to developing a disorder if exposed to stressful or traumatic experiences (Davis). Despite advances in psychiatric genetics, identifying specific genes linked to anxiety is still difficult due to small sample sizes, overlapping

symptoms, and high rates of multiple disorders occurring together (Purves et al.). Current research estimates the heritability of anxiety to be around 25-60%, and this is consistent across children, teenagers, and adults (Purves et al.)

Other biological markers, such as microRNAs found in blood, are also connected to anxiety (Łoś and Waszkiewicz). These markers are associated with brain-derived neurotrophic factor (BDNF), a protein that helps brain cells grow, stay healthy, and connect with each other, as well as to regulate stress by the hypothalamic-pituitary-adrenal (HPA) axis (Łoś and Waszkiewicz). Biomarkers may reflect increased or decreased activity, depending on the system and, with time, may help determine who is at risk and lead to a more personalized approach to treatment (Łoś and Waszkiewicz). Studies also show that parental anxiety or depression significantly increases the risk of adolescent anxiety which suggests both a shared genetic risk and environmental influences (Narmandakh et al.). Interestingly, the mix of genetic and environmental risk factors look very similar for both men and women (Hettema et al.).

Anxiety is believed to be multidetermined, reflecting influences from biological, psychological, and social factors, suggesting that no single biomarker would probably be able to explain the disorder alone (Łoś and Waszkiewicz). Anxiety disorders are associated with a combination of biological, psychological, and social factors that interact with environmental and lifestyle aspects including stress, physical exercise, comorbid health problems and drugs (Łoś & Waszkiewicz). Biomarkers may also link to clusters of symptoms, rather than to one clear diagnosis, which shows just how complex anxiety really is (Łoś and Waszkiewicz). An example of this is shown in a study by Narmandakh et al., which utilized this biopsychosocial approach, found that the female sex was the most robust and consistent predictor for anxiety. It also showed that temperamental traits like high frustration and low self-control increased the risk, and that parental psychopathology, particularly anxiety or depression, raised adolescent risk (Narmandakh et al.).

Anxiety is also related to how the brain processes fear. Davis explains that people show increased patterns of neural activity towards threats even in the absence of actual danger. He notes that this hypersensitivity causes constant overthinking, high alertness, and difficulty calming down. Simply the anticipation of a stressful event can cause an intense physical response, for example a racing heart, sweating or stomach discomfort, before anything even happens. This raises the possibility that the brain may react to what could happen as opposed to what is happening. Davis further points out that a key player is the amygdala – the part of your brain that manages fear – which is overactive in people with anxiety. This overactivity can produce fear signals even in mildly threatening environments, resulting in emotional and physical symptoms (Davis).

Structural changes in the amygdala and hippocampus are also promising markers of anxiety. For example, the amygdala is often enlarged or hyperactive, contributing to heightened fear responses, while the hippocampus frequently shows reduced volume, weakening its ability to regulate stress and distinguish between safe and threatening contexts (Łoś and Waszkiewicz). Cerebrospinal fluid (CSF) biomarkers also reveal shifts: lower oxytocin and 5-HIAA (a serotonin

metabolite) are correlated to higher levels of anxiety (Łoś and Waszkiewicz). Early stress may also continue to sculpt neurodevelopment in areas such as the amygdala and prefrontal cortex (Narmandakh et al.). Neuroimaging findings further support these results by establishing consistent structural differences in both the amygdala and the hippocampus, as well as evidence of local brain inflammation and microglial activation that may reflect the underlying biology of anxiety (Łoś and Waszkiewicz).

Recent studies have also focused on the gut-brain axis (Malan-Muller et al.). Stress responses can be measured through saliva markers like cortisol, lysozyme, and alpha-amylase, while blood biomarkers like serotonin and inflammatory markers reveal how gut health and anxiety are linked (Malan-Muller et al.). The microbiota-gut brain (MGB) axis is seen as a major system in the development of anxiety and how people respond to treatment. This study by Malan-Muller et al. highlights that when the gut microbiome is balanced, it supports healthy stress responses and produces neurotransmitters that regulate mood. But when there is an imbalance – such as too much harmful bacteria or not enough beneficial bacteria – communication between the gut and brain can become disrupted (Malan-Muller et al.). They later note that early in life, the gut microbiome helps shape the HPA axis that regulates stress responses, meaning that gut health can establish long-term sensitivity to anxiety. Even in adulthood, gut imbalances can increase anxiety by lowering protective bacteria, altering neurotransmitter production, or rising inflammation. Stress itself can also disrupt gut bacteria, increasing gut permeability and allowing bacterial byproducts to reach the brain, which could impact its function and raise the risk of anxiety. The gut microbiome can also form neurotransmitters like GABA and serotonin that then enter the bloodstream and affect the brain, both of which are associated with mood and anxiety regulation (Malan-Muller et al.).

Altogether, these findings show that biological factors including genetic vulnerability, brain chemistry and neural activity, neurophysiological and structural changes, biomarkers and biological indicators, and the gut-brain axis, all create a solid foundation for anxiety disorders. However, these biological vulnerabilities do not act in isolation. They interact with psychological and social factors, such as early life stress, personality traits, and environment experiences, to shape the onset, severity, and type of anxiety an individual may develop. Understanding these complex interactions not only highlights the multidimensional nature of anxiety disorders but also reinforces the importance of personalized approaches for prevention and treatment. By considering the bio-psycho-social influences, researchers and doctors can more accurately detect individuals at-risk, tailor interventions, and ultimately improve outcomes for those affected by anxiety.

Psychological Factors Contributing to Anxiety Disorders

Psychological experiences and traits are crucial in shaping vulnerability to anxiety disorders, often determining whether biological predispositions progress into clinical conditions. Among these, early-life stress is one of the strongest predictors. Experiences such as emotional abuse, bullying, neglect, or low socioeconomic status significantly increase the risk of adult

anxiety, with emotional trauma emerging as the most consistent factor (Lähdepuro et al.) These effects accumulate, with individuals exposed to several forms of trauma are much more likely than others to develop clinically intense levels of anxiety (Lähdepuro et al.). These adverse experiences disrupt the development of brain regions that regulate emotion, including the amygdala (heightened fear reactivity), hippocampus (weakened memory regulation of stress), and frontal cortex (impaired reasoning and inhibition), and alter the hypothalamic-pituitary-adrenal (HPA) axis, which makes individuals hypersensitive to stress (Juruena et al.). They also dysregulate the HPA axis that regulates stress hormones such as cortisol, rendering them overly sensitive to even mild stressors (Juruena et al.).

In adulthood, chronic stress is a key driver of anxiety. Ongoing pressures such as work overload, trauma, financial insecurity and relational strain can maintain high arousal levels, specifically in people who have experienced childhood trauma (Hussenoeder et al.) For example, factors including emotional trauma like physical, sexual or emotional abuse have been associated with greater risk of developing anxiety disorders in adulthood, with different trauma types associated with higher odds of developing specific conditions (Fernandes and Osório). Emotional trauma is, for instance, most associated with social anxiety and physical trauma more related to panic disorder, likely because of negative cognitive schemas which is a mental framework for interpreting the world. (Fernandes and Osório; Kalin). Kalin explains that survivors of emotional abuse frequently internalize beliefs of unworthiness or are in a constant state of performance anxiety that inevitably results in chronic, adult-onset anxiety. Furthermore, trauma may induce permanent epigenetic changes in stress-related genes that confer vulnerability throughout the lifespan (Lähdepuro et al.). For instance, Lähdepuro et al. explored the way childhood adversity affects gene expression and found that stress can ‘switch on’ genes that heighten anxiety reactivity – explaining why early trauma leaves a biological scar.

Individual differences in traits of temperament and personality are also strong predictors of risk for anxiety. Children who are highly shy, behaviorally inhibited, or are easily frustrated are associated with increased vulnerability particularly in the absence of effortful control – the ability to regulate attention and impulses (Narmandakh et al.; Murray et al.). Effortful control involves skills such as shifting attention away from threatening thoughts or pausing before reacting, and Murray et al.’s longitudinal study followed children over several years to show that those who struggled most with this regulation were significantly more likely to develop anxiety symptoms later in adolescence (Narmandakh et al.; Murray et al.). Though these characteristics may sound like common variants of normal behavior, they can escalate into anxiety disorders when combined with environmental stressors, including overprotective parenting or peer rejection. Conversely, strong self-regulation and attentional control act as protective factors, enabling adaptive responses to stress.

As these early temperamental traits develop over time, broader personality traits also shape how individuals respond to stress and anxiety in adulthood. The Big Five Model, as described by Tuygar Okutucu and Ceyhun, identifies five broad traits: neuroticism, extraversion, conscientiousness, agreeableness, and openness. They note that higher levels of neuroticism are

associated with enhanced anxiety, and lower levels of extraversion, conscientiousness, agreeableness and openness have all been tied to greater vulnerability. For example, individuals with high neuroticism tend to interpret situations as more threatening, which amplifies social stressors such as peer rejection or cultural expectations (Tuygar Okutucu and Ceyhun). On the other side, traits such as extraversion and conscientiousness can act as protective factors by supporting stronger social networks and more effective coping mechanisms (Tuygar Okutucu and Ceyhun). Tuygar Okutucu and Ceyhun also observe that anxiety not only correlates with personality but can influence it – prolonged anxiety may heighten neuroticism or reduce openness over time, especially in individuals exposed to chronic stress. Positive environments, however, can buffer these effects, fostering resilience, and encouraging adaptive personality changes (Tuygar Okutucu and Ceyhun).

Parental influences further shape these outcomes. Children often model their parents' emotional responses, meaning anxious or overcontrolling parenting styles can inadvertently teach avoidance behaviors (Murray et al.). Parents' anxiety and depression not only model maladaptive coping, but they may also offer less emotional support, which can inadvertently teach avoidance behaviors (Murray et al.). Murray et al. showed this by observing parent-child interactions and found out that children of anxious parents were more likely to exhibit avoidance behaviors, suggesting that parental influence is not only genetic but also learned. This interaction between parental psychopathology and child temperament compounds vulnerability, transmitting risk across generations.

Taken together, these findings show that psychological factors are central in shaping anxieties risk, interacting dynamically with biological vulnerabilities to determine both the onset and severity of anxiety disorders. Early adversity alters stress systems and cognitive patterns, while temperament and parenting shape coping styles. Overtime, personality, and chronic stress in adulthood sustains symptoms. Engaging with such studies, it becomes apparent that the researchers are not just uncovering risk factors but are pulling back a curtain to show the mechanisms – from gene expression to parenting styles – that lead some people to develop anxiety disorders when others remain unaffected. These psychological foundations also interact with social contexts, meaning that traits like neuroticism can amplify the impact of family, peer, or cultural stressors.

Social Factors Contributing to Anxiety Disorders

Social influences such as family dynamics, cultural expectations, peer relationships, and societal pressures can contribute to the onset and intensity of anxiety disorders. These influences shape how individuals perceive themselves and the world, often reinforcing patterns of fear and worry. Biology may underpin the basic vulnerability, but social contexts can either buffer against or heighten anxiety symptoms depending on the quality of the relationships and pressures individuals face.

One of the strongest social predictors of anxiety is family environment. Early family

relationships matter because they influence children's ability to regulate their emotions and cope. For instance, research in attachment theory has shown that erratic and over-control caregiving can lead to insecurity and hypervigilance in children, which can serve as a base for subsequent anxiety (Möller et al.). This is believed to occur because children are taught to anticipate uncertainty or danger in their environment, and as they age, they become hyper vigilant for stressors (Möller et al.). These findings are supported by longitudinal studies, though much of the data relies on retrospective accounts, which creates some margin for recall bias (Möller et al.).

Parenting styles also contribute to the development of anxiety by influencing children's sense of autonomy. Some overprotective or authoritarian parents may have the best interests of their children at heart, but they behave in such a way that communicates that the world is unsafe and the child is incapable of managing it – this encourages learned dependency and avoidance, which are two core features of anxiety disorders (McLeod et al.). Conversely, parents who balance support with autonomy encourage resilience and self-confidence, which protects children from the risks of anxiety development (McLeod et al.). An important takeaway is that parenting-style research varies across cultural contexts and that “overprotective” in reference to a Western population, could be seen as normal or caring in collectivist cultures (Varela and Hensley-Maloney). This implies that researchers need to make tentative interpretations and attempt to explore the reasons why those distinct parenting practices may emerge.

Beyond the family, peer relationships and school environments are potent social factors. Adolescence, in particular, is a period where fear of social evaluation peaks, which makes peer rejection or bullying especially harmful. While direct evidence from school-aged populations is limited, Barnoa and Ward examined socially and psychologically marginalized adults in forensic inpatient settings. Their findings suggest that experiences of stigma, rejection and alienation exacerbate psychological problems and discourage recovery. Using a literature review and small case studies, they found that when people's sense of belonging and identity were supported, motivation and rehabilitation outcomes had improved. Although the research was limited due to a limited number of 5 participants and that long-term outcomes were not measured. This leaves questions about the precise mechanisms by which social rejection, including bullying from peers in adolescence, led to continued anxiety disorders into adulthood. However, their model still reveals and underscores that rejection is often internalized as a personal flaw, which fuels anxious thoughts and reinforces vulnerability to mental health disorders (Barnoa and Ward).

Cultural expectations also influence anxiety by placing pressure on individuals to meet strict standards. In collectivist societies, anxiety may develop from fears of bringing shame to the family or community, whereas in individualist cultures, pressure to achieve independence and success can create its own stress (Sells et al.) For instance, in East Asian societies, high academic pressure has been associated with anxiety because 'failure' is seen not only as a reflection of personal inadequacy but also considered as a loss of family honor (Biggart et al.). This may mean that cultural values, rather than promoting solidarity and motivation, have unintended effects by intensifying perfectionism and increasing fear of failure (Biggart et al.). However, one limitation in such studies is determining whether the influence is really cultural or is also related

to other factors – like socioeconomic status or expectations about parenting – can be difficult, and so a clear chain of causation doesn't always emerge.

Societal pressures, particularly those tied to media and socioeconomic environment, also play a role in shaping anxiety. For example, exposure to idealized body images in social media can heighten self-comparison and dissatisfaction which is closely related with anxiety among teenagers (Kapoor). Similarly, living under chronic economic strain or in unstable communities creates a constant threat and fear on individuals, which activates stress symptoms in the body and heightens vulnerability to anxiety (Kessler et al.). The reason these pressures are so influential is because they threaten two fundamental needs: the need for stability and self-worth. When people feel they are unable to meet social expectations or control the environment, their anxiety increases (Kessler et al.).

Overall, social factors contribute to anxiety disorders by shaping how individuals interpret their environment and their ability to cope with challenges. From family attachments to peer relationships, cultural pressures, and societal influences, these factors do not just expose individuals to stress – they also teach them how to think about stress. The primary source of error in this area of research is the inability to disentangle overlapping social variables which makes it difficult to determine whether these factors contribute alongside anxiety or are exclusively associated with it. Still, the recurring evidence highlights the need to view anxiety not only as an internal disorder but also as a reflection of the broader social contexts in which individuals live (Church et al.).

Conclusion

This review aimed to identify the most significant biological, psychological, and social factors contributing to the development of anxiety disorders. Across studies, it is clear that anxiety results from a complex interplay between inherited vulnerabilities, cognitive patterns, and environmental influences, supporting the biopsychosocial model.

Biologically, research indicate genetic predispositions, neurotransmitter imbalances, structural and functional differences in the amygdala and hippocampus, dysregulation of the hypothalamic-pituitary-gland (HPA) axis, and gut-brain axis imbalances significantly contribute to heightened susceptibility to anxiety (Hettema et al.; Murray et al.; Hussenoder et al.; Kalin). These factors create a foundation that can be amplified by psychological and social experiences.

Psychological factors, including early-life trauma, maladaptive cognitive patterns, temperament, personality traits such as high neuroticism, and chronic stress in adulthood, further shape vulnerability (Fernandes and Osório; Kalin; Hussenoder et al.; Tuygar Okutucu and Ceyhun). These factors interact with biological predispositions, demonstrating how trauma and maladaptive thinking can translate genetic or neurobiological risk into clinical anxiety (Fernandes and Osório; Kalin).

Social influences, including family environment, parenting styles, peer relationships, cultural expectations, and societal pressures, also play a pivotal role in shaping both risk and resilience (Barnao and Ward; Möller et al.; Varela and Hensley-Maloney; Sells et al.; Biggart et al.). These contributors interact with biological vulnerabilities to illustrate how early trauma and

maladaptive cognition can transform genetic or neurobiological vulnerability into a clinical anxiety syndrome. Together, these results show that no single cause can be attributed for anxiety and underscore the importance of considering multiple interacting factors.

Despite the amount of work, gaps remain. Many studies focus on isolated factors rather than analyzing the dynamic interplay of the biological, psychological, and social variables simultaneously. Future studies may foster a deep phenotyping approach, integrating genetic, neuroimaging, physiological, and behavioral data to gain better insights into the mechanisms of anxiety and associated with heterogeneity in symptom presentation (Wright and Herzberg). Longitudinal studies, which examine how early-life experiences interact with biology over time are also scarce. Furthermore, social and cultural impacts are evident, but research is limited in its diversity or with regards to intersectional impact.

Taken together, this review shows that anxiety disorders are best understood through a biopsychosocial lens. Biological vulnerabilities lay the groundwork, psychological experience determines an individual's trajectory, and social context influences the severity and persistence of symptoms. By integrating these perspectives and addressing current research gaps, future studies can develop more personalized, effective prevention and treatment strategies, ultimately improving outcomes for people living with anxiety.

Works Cited

- Barnao, Mary, and Tony Ward. "Sailing Uncharted Seas without a Compass: A Review of Interventions in Forensic Mental Health." *Aggression and Violent Behavior*, vol. 22, May 2015, pp. 77–86. DOI.org (Crossref), <https://doi.org/10.1016/j.avb.2015.04.009>.
- Biggart, Laura, et al. "The Team as a Secure Base: Promoting Resilience and Competence in Child and Family Social Work." *Children and Youth Services Review*, vol. 83, Dec. 2017, pp. 119–30. DOI.org (Crossref), <https://doi.org/10.1016/j.chidyouth.2017.10.031>.
- Buckner, Julia D., et al. "Biopsychosocial Model Social Anxiety and Substance Use Revised." *Current Psychiatry Reports*, vol. 23, no. 6, Jun. 2021, p. 35. DOI.org (Crossref), <https://doi.org/10.1007/s11920-021-01249-5>.
- Church, Nicole F., et al. "Do Dysfunctional Cognitions Mediate the Relationship between Risk Factors and Postnatal Depression Symptomatology?" *Journal of Affective Disorders*, vol. 87, no. 1, Jul. 2005, pp. 65–72. DOI.org (Crossref), <https://doi.org/10.1016/j.jad.2005.03.009>.
- Davis, M. "The Role of the Amygdala in Fear and Anxiety." *Annual Review of Neuroscience*, vol. 15, no. 1, Mar. 1992, pp. 353–75. DOI.org (Crossref), <https://doi.org/10.1146/annurev.ne.15.030192.002033>.
- Domschke, Katharina, and Eduard Maron. "Genetic Factors in Anxiety Disorders." *Modern Trends in Psychiatry*, edited by D.S. Baldwin and B.E. Leonard, vol. 29, S. Karger AG, 2013, pp. 24–46. DOI.org (Crossref), <https://doi.org/10.1159/000351932>.
- Fernandes, V., and F.L. Osório. "Are There Associations Between Early Emotional Trauma and Anxiety Disorders? Evidence from a Systematic Literature Review and Meta-Analysis." *European Psychiatry*, vol. 30, no. 6, Sep. 2015, pp. 756–64. DOI.org (Crossref), <https://doi.org/10.1016/j.eurpsy.2015.06.004>.
- Hettema, John M., Michael C. Neale, et al. "A Review and Meta-Analysis of the Genetic Epidemiology of Anxiety Disorders." *American Journal of Psychiatry*, vol. 158, no. 10, Oct. 2001, pp. 1568–78. DOI.org (Crossref), <https://doi.org/10.1176/appi.ajp.158.10.1568>.
- Hettema, John M., Carol A. Prescott, et al. "The Structure of Genetic and Environmental Risk Factors for Anxiety Disorders in Men and Women." *Archives of General Psychiatry*, vol. 62, no. 2, Feb. 2005, pp. 182–89, <https://doi.org/10.1001/archpsyc.62.2.182>.
- Hussenoeder, Felix S., et al. "Connecting Chronic Stress and Anxiety: A Multi-Dimensional Perspective." *Psychology, Health & Medicine*, vol. 29, no. 3, Mar. 2024, pp. 427–41. DOI.org (Crossref), <https://doi.org/10.1080/13548506.2022.2124292>.
- Jenkins, Paul E., et al. "Eating Disorders and Quality of Life: A Review of the Literature." *Clinical Psychology Review*, vol. 31, no. 1, Feb. 2011, pp. 113–21. DOI.org (Crossref), <https://doi.org/10.1016/j.cpr.2010.08.003>.
- Juruena, Mario F., et al. "The Role of Early Life Stress in HPA Axis and Anxiety." *Anxiety Disorders*, edited by Yong-Ku Kim, vol. 1191, Springer Singapore, 2020, pp. 141–53.

- Advances in Experimental Medicine and Biology. DOI.org (Crossref), https://doi.org/10.1007/978-981-32-9705-0_9.
- Kalin, Ned H. "The Critical Relationship Between Anxiety and Depression." *American Journal of Psychiatry*, vol. 177, no. 5, May 2020, pp. 365–67. DOI.org (Crossref), <https://doi.org/10.1176/appi.ajp.2020.20030305>.
- Kapoor, Isha. "Social Anxiety Disorder Among Adolescents in Relation to Peer Pressure and Family Environment." *Bioscience Biotechnology Research Communications*, vol. 13, no. 2, Jun. 2020, pp. 923–29. DOI.org (Crossref), <https://doi.org/10.21786/bbrc/13.2/80>.
- Kessler, Ronald C., et al. "Prevalence, Severity, and Comorbidity of 12-Month DSM-IV Disorders in the National Comorbidity Survey Replication." *Archives of General Psychiatry*, vol. 62, no. 6, Jun. 2005, p. 617. DOI.org (Crossref), <https://doi.org/10.1001/archpsyc.62.6.617>.
- Kessler, Ronald C., and Paul E. Greenberg. "The economic burden of anxiety and stress disorders." *Neuropsychopharmacology: The fifth generation of progress* 67 (2002): 982-92
- Lähdepuro, Anna, et al. "The Impact of Early Life Stress on Anxiety Symptoms in Late Adulthood." *Scientific Reports*, vol. 9, no. 1, Mar. 2019, p. 4395. DOI.org (Crossref), <https://doi.org/10.1038/s41598-019-40698-0>.
- Łoś, Kacper, and Napoleon Waszkiewicz. "Biological Markers in Anxiety Disorders." *Journal of Clinical Medicine*, vol. 10, no. 8, Apr. 2021, p. 1744. DOI.org (Crossref), <https://doi.org/10.3390/jcm10081744>.
- Malan-Muller, Stefanie, et al. "The Gut Microbiome and Mental Health: Implications for Anxiety- and Trauma-Related Disorders." *OMICS: A Journal of Integrative Biology*, vol. 22, no. 2, Feb. 2018, pp. 90–107. DOI.org (Crossref), <https://doi.org/10.1089/omi.2017.0077>.
- McKay, Michael T., et al. "Childhood Trauma and Adult Mental Disorder: A Systematic Review and Meta-analysis of Longitudinal Cohort Studies." *Acta Psychiatrica Scandinavica*, vol. 143, no. 3, Mar. 2021, pp. 189–205. DOI.org (Crossref), <https://doi.org/10.1111/acps.13268>.
- McLeod, Bryce D., et al. "Examining the Association between Parenting and Childhood Anxiety: A Meta-Analysis." *Clinical Psychology Review*, vol. 27, no. 2, Mar. 2007, pp. 155–72. DOI.org (Crossref), <https://doi.org/10.1016/j.cpr.2006.09.002>.
- Möller, Eline L., et al. "Associations between Maternal and Paternal Parenting Behaviors, Anxiety and Its Precursors in Early Childhood: A Meta-Analysis." *Clinical Psychology Review*, vol. 45, Apr. 2016, pp. 17–33. DOI.org (Crossref), <https://doi.org/10.1016/j.cpr.2016.03.002>.
- Narmandakh, Altanzul, et al. "Psychosocial and Biological Risk Factors of Anxiety Disorders in Adolescents: A TRAILS Report." *European Child & Adolescent Psychiatry*, vol. 30, no. 12, Dec. 2021, pp. 1969–82. DOI.org (Crossref), <https://doi.org/10.1007/s00787-020-01669-3>.

- Purves, Kirstin L., et al. "A Major Role for Common Genetic Variation in Anxiety Disorders." *Molecular Psychiatry*, vol. 25, no. 12, Dec. 2020, pp. 3292–303. DOI.org (Crossref), <https://doi.org/10.1038/s41380-019-0559-1>.
- Ryu, Soomin, and Lu Fan. "The Relationship Between Financial Worries and Psychological Distress Among U.S. Adults." *Journal of Family and Economic Issues*, vol. 44, no. 1, Mar. 2023, pp. 16–33. DOI.org (Crossref), <https://doi.org/10.1007/s10834-022-09820-9>.
- Sells, James N., et al. "Relationship of Supervisor and Trainee Gender to In-Session Verbal Behavior and Ratings of Trainee Skills." *Journal of Counseling Psychology*, vol. 44, no. 4, Oct. 1997, pp. 406–12. DOI.org (Crossref), <https://doi.org/10.1037/0022-0167.44.4.406>.
- Tuygar Okutucu, Fatma, and Hacer Akgul Ceyhun. "The Impact of Anxiety and Depression Levels on the Big Five Personality Traits." *PLOS One*, edited by Frantisek Sudzina, vol. 20, no. 7, Jul. 2025, p. e0321373. DOI.org (Crossref), <https://doi.org/10.1371/journal.pone.0321373>.
- Varela, R. Enrique, and Lauren Hensley-Maloney. "The Influence of Culture on Anxiety in Latino Youth: A Review." *Clinical Child and Family Psychology Review*, vol. 12, no. 3, Sep. 2009, pp. 217–33. DOI.org (Crossref), <https://doi.org/10.1007/s10567-009-0044-5>.
- Wright, J.T., and M.C. Herzberg. "Science for the Next Century: Deep Phenotyping." *Journal of Dental Research*, vol. 100, no. 8, Jul. 2021, pp. 785–89. DOI.org (Crossref), <https://doi.org/10.1177/00220345211001850>.

Improvement of GraphCast Temperature Field and Prediction of Mid-Latitude Fronts Based on PINN Hard Constraints By Taoyu Pan

Abstract

This study presents a comprehensive framework to enhance the GraphCast model's temperature field predictions and mid-latitude frontal analyses by integrating physics-informed neural networks (PINNs) with hard constraints. By embedding fundamental atmospheric physical laws—thermal wind balance, mass conservation, and vorticity advection—into the training process, the model ensures strict adherence to meteorological principles while leveraging GraphCast's graph neural network (GNN) for spatial-temporal pattern recognition. Quantitative evaluations on ERA5 reanalysis data demonstrate a 15–20% reduction in temperature root mean square error (RMSE) and a 25–30% improvement in frontal detection F1-scores compared to the baseline GraphCast. The framework bridges data-driven efficiency with physics-based rigor, offering a novel approach to operational weather forecasting. This work also explores the decomposition of GraphCast into sub-modules (e.g., temperature, wind) to isolate and enforce hard constraints on underperforming components, as suggested by domain experts.

Keywords

GraphCast Model Enhancement; Physics-Informed Neural Networks; Hard Physical Constraints; Mid-Latitude Frontal Analysis

1. Introduction

1.1 Background and Context

The GraphCast model, a graph neural network (GNN)-based system developed by Google Research, has demonstrated remarkable accuracy in medium-range weather prediction, outperforming traditional numerical weather prediction (NWP) models in certain metrics (Lam et al., 2023). However, like many data-driven approaches, GraphCast relies heavily on empirical loss functions (e.g., mean squared error) that do not explicitly enforce physical consistency. This limitation becomes critical when predicting phenomena governed by well-understood atmospheric dynamics, such as mid-latitude fronts, where violations of conservation laws (e.g., mass, momentum) can lead to unrealistic forecasts.

Mid-latitude fronts, characterized by sharp temperature gradients and vorticity advection, are pivotal in storm formation and precipitation patterns. Traditional frontal analysis relies on thermodynamic criteria (e.g., the 2 K/100 km temperature gradient threshold) and dynamic criteria (e.g., positive vorticity advection). While GraphCast captures broad frontal patterns, its data-driven nature occasionally produces physically inconsistent results, such as discontinuous temperature fields or unrealistic frontal propagation speeds.

1.2 Research Objectives

This study aims to address these limitations by:

1. Enhancing Temperature Field Accuracy: Enforcing thermal wind balance and mass conservation as hard constraints to ensure temperature predictions align with vertical wind shear and air density dynamics.

2. Improving Frontal Detection: Integrating vorticity advection criteria into the PINN framework to guide mid-latitude frontal localization.

3. Validating Robustness: Comparing performance against operational models (e.g., ECMWF) and ablating constraint impacts to quantify contributions.

4. Decoupling GraphCast: Isolating temperature-related components to apply targeted hard constraints, as suggested by domain experts.

2. Theoretical Foundation

2.1 Atmospheric Physical Laws

2.1.1 Thermal Wind Balance

The thermal wind equation is a cornerstone of atmospheric dynamics, linking horizontal temperature gradients to vertical wind shear. Derived from the primitive equations under the assumptions of geostrophic and hydrostatic balance, it provides a critical relationship for validating temperature fields in numerical models. The equation is expressed as:

where $\mathbf{u}$ is the horizontal wind vector, g is gravitational acceleration, f_0 is the Coriolis parameter, T_0 is a reference temperature, and ∇T is the horizontal temperature gradient. This equation ensures that temperature variations drive wind shear, a critical relationship for maintaining physical consistency in temperature fields.

In mid-latitudes, the thermal wind equation explains the west-to-east orientation of jet streams, where sharp temperature contrasts between polar and tropical air masses generate strong wind shear.

A positive (negative) temperature gradient in the north-south direction induces a westerly (easterly) vertical wind shear, aligning with the observed strengthening of westerlies with height in warm-core systems.

In our framework, the thermal wind equation is embedded as a hard constraint to ensure temperature predictions align with observed wind shear patterns. By penalizing deviations from this relationship, the model avoids unphysical temperature fields that could arise from pure data-driven training, such as discontinuous gradients or unrealistic jet stream orientations.

2.1.2 Mass Conservation

The continuity equation governs the conservation of mass in the atmosphere. For an incompressible fluid (a valid approximation for synoptic-scale motions where density variations are small), it simplifies to:

$$\nabla \cdot \mathbf{u} = 0$$

This equation states that the divergence of the horizontal wind field must balance vertical motion, ensuring no net creation or destruction of air mass within a control volume.

2.1.3 Vorticity Advection

Frontal dynamics are influenced by vorticity advection:

where ζ is relative vorticity, f is the planetary vorticity, and $\mathbf{V}$ is the wind vector. Positive vorticity advection (PVA) ahead of a trough is a key indicator of mid-latitude cyclone development.

2.2 Physics-Informed Neural Networks (PINNs)

Physics-Informed Neural Networks (PINNs) represent a paradigm shift in machine learning by explicitly integrating governing physical laws into the training process. Unlike conventional neural networks that rely solely on data-driven loss functions (e.g., mean squared error), PINNs augment their objective functions with terms derived from partial differential equations (PDEs), ordinary differential equations (ODEs), or algebraic constraints. This hybrid approach ensures that predictions adhere to known physical principles, improving robustness in scenarios where data is scarce or noisy.

The enforcement of physical laws within Physics-Informed Neural Networks (PINNs) represents a fundamental design choice with profound implications for model accuracy and reliability. This distinction between soft and hard constraints shapes how neural networks interact with domain-specific knowledge, balancing computational efficiency against physical fidelity. Soft constraints operate through penalty-based approaches, where physical residuals are incorporated as additive terms in the loss function. These residuals, derived from governing equations such as the Navier-Stokes equations or elasticity theory, quantify deviations from expected physical behavior. By weighting these terms with a hyperparameter λ , practitioners can adjust the influence of physical principles relative to data fidelity. This flexibility makes soft constraints straightforward to implement using standard deep learning frameworks like PyTorch or TensorFlow, as they leverage automatic differentiation for gradient computation. However, this approach introduces challenges in tuning λ to avoid under- or over-constraining the model. In low-data regimes, soft constraints may struggle to enforce physical consistency, as the network prioritizes minimizing data-driven errors over satisfying theoretical relationships.

Recent advancements have significantly expanded the applicability of hard constraints. In aerodynamics, a PINN enforcing Euler equations achieved sub-2% error in airfoil pressure predictions by embedding inviscid flow constraints directly into the network layers. This outperformed both soft-constrained PINNs (15% error) and traditional CFD solvers (5% error) while maintaining computational efficiency. Structural mechanics applications demonstrate similar benefits: Ouyang et al.'s convolutional neural network for glass pane analysis reduced spurious deformations by 90% by embedding linear elasticity equations, ensuring displacement

fields satisfy $\nabla \cdot \sigma + F = 0$. Fracture mechanics models now use Griffith's criterion as a hard constraint to guide crack propagation paths, aligning predictions with experimental observations of material failure.

The earthquake engineering case study highlights these advantages in operational contexts. A PINN constrained by the momentum balance equation predicted building responses during the 1994 Northridge earthquake with 10% error, a threefold improvement over soft-constrained models. This accuracy stems from the network's ability to enforce equilibrium conditions at every time step, preventing unphysical accelerations that plague data-only approaches. Such results underscore the value of hard constraints in safety-critical applications where physical plausibility is non-negotiable.

3. Methodology

3.1 GraphCast Optimization

In the optimization process of GraphCast, the research team made multidimensional improvements to address the limitations of the baseline architecture, focusing on graph structure optimization, physical constraint integration, and training efficiency improvement. Firstly, when the baseline encoder maps the latitude and longitude grids to multi-grid nodes, it adopts a six-iteration layering strategy of icosahedral grids to generate a multi-scale grid set with gradually increasing resolution. This hierarchical structure reduces the number of global nodes from the original 1,038,240 to 40962. The nested relationship of coarse and fine grid nodes allows long-distance edges to complete information exchange within a few message transmission steps while maintaining spatial uniformity. For example, the 16-layer deep map attention network (GAT) in the processor module dynamically assigns weights to enable each node to adaptively adjust its attention intensity to different meteorological elements (such as temperature and wind field) when aggregating neighbor information. This multi-head attention mechanism effectively captures key patterns in atmospheric circulation, such as the interaction between the core area of the jet stream and the high-pressure system.

To address the issue of insufficient physical consistency in the baseline model, the optimized version introduces a hybrid mechanism that combines soft and hard constraints. In the encoding stage, when mapping the original meteorological variables to the latent space through an embedder, an additional physical residual term is added as a regularization condition. Specifically, the law of conservation of mass is transformed into a constraint on the difference in water vapor flux over continuous time steps, while the law of conservation of momentum is soft-penalized through the equilibrium equation of pressure gradient force and Coriolis force. This constraint is manifested as an auxiliary term in the loss function, which, together with the prediction error, forms the multi-task learning objective. The weight coefficients are set to 0.15 after hyperparameter tuning to ensure a balance between physical fidelity and prediction accuracy. It is worth noting that the hard-constrained version, through architectural modifications, directly embeds the discretized form of partial differential equations (PDEs) in the

message passing process. For example, each GAT layer of the processor layer forces nodes to update the spatial derivative relationship that satisfies the advection equation. This modification enables the model to maintain end-to-end differentiability while significantly reducing non-physical oscillation phenomena in prediction.

In terms of training strategy, the research team adopts a progressive curriculum learning framework, initially using only low-resolution (1.4 °) ERA5 data for pre-training, gradually transitioning to high-resolution data at 0.25 °. This strategy avoids the gradient vanishing problem caused by direct high-resolution training while utilizing the fast iteration characteristics of low-resolution data to accelerate the model's capture of basic atmospheric dynamics patterns. In distributed training, 32 Cloud TPU v4 devices increased the batch size to 2,048 through a hybrid strategy of data parallelism and model parallelism. Combined with gradient accumulation technology, the effective data volume for a single parameter update was expanded to 8,192 samples while maintaining memory efficiency. In addition, the optimizer adopts the AdamW variant, with an initial learning rate set at 2.5e-4 and a cosine annealing strategy, gradually decreasing to 1e-5 in the last 20% of the training cycle. This dynamic adjustment effectively alleviates the problem of error accumulation in the long-term prediction of the model.

3.1.1 Baseline Architecture

GraphCast's architecture is a graph neural network (GNN) designed to model spatial-temporal dependencies in atmospheric data. The encoder processes gridded meteorological fields into latent graph representations. Each grid cell is treated as a node in a graph, with edges dynamically defined by geographic proximity and meteorological similarity. The encoder uses graph convolution layers to aggregate information from neighboring nodes, transforming raw data into a latent space of dimension $\mathbb{R}^{N \times D}$, where N is the number of nodes (e.g., 10,368 for 0.25° resolution), and D is the hidden dimension (e.g., 256).

The processor applies message-passing across graph nodes to capture interactions between atmospheric variables. It consists of L layers of graph attention networks (GATs), where each layer updates node features by aggregating messages from neighbors. The attention mechanism weights contributions based on feature similarity, allowing the model to focus on relevant spatial relationships (e.g., fronts, jets). The processor's output is a refined latent representation that encodes multi-scale atmospheric dynamics.

The decoder projects the latent features back to gridded outputs, reconstructing the predicted fields at each node. It uses transpose graph convolutions to upsample the latent space to the original grid resolution, ensuring spatial consistency with the input data.

Input/Output Specifications:

Input: 6-hourly ERA5 reanalysis data (temperature, wind components, pressure) at 0.25° resolution.

Output: 6-hourly forecasts of the same variables up to 240 hours ahead.

3.1.2 Multi-Scale Feature Extraction

The integration of multi-scale feature extraction within GraphCast's graph convolution layers represents a critical advancement in resolving fine-scale temperature gradients essential for accurate mid-latitude frontal analysis. Traditional graph neural networks often struggle to capture spatial hierarchies in meteorological data, particularly when dealing with phenomena that exhibit variability across different scales. Frontal systems, for instance, manifest as sharp temperature contrasts at regional levels while being influenced by larger-scale atmospheric waves such as Rossby waves. This duality necessitates a feature extraction mechanism capable of simultaneously resolving micro-scale frontal boundaries and macro-scale dynamical patterns.

The adaptive graph convolution approach addresses this challenge through dynamic edge weighting mechanisms that prioritize nodes exhibiting significant meteorological activity. By identifying regions where temperature gradients exceed 1.5 K/100 km, the model assigns higher connectivity weights to nodes within these frontal zones. This ensures that interactions between adjacent nodes with strong thermal contrasts are emphasized during message passing. The self-attention mechanism further refines this connectivity by computing edge weights based on the similarity of node features. Query and key vectors derived from local atmospheric dynamics enable the model to establish stronger connections between nodes with converging wind patterns or similar pressure tendencies, effectively creating a dynamically adjusted graph topology that mirrors the physical relationships governing atmospheric flow.

Complementing this adaptive connectivity scheme is the multi-scale kernel fusion strategy. Conventional graph convolutions often employ fixed kernel sizes, limiting their ability to capture phenomena across different spatial scales. By combining 3×3 , 5×5 , and 7×7 kernels, the model gains the capacity to resolve both small-scale features, like thunderstorm complexes, and synoptic-scale systems, such as troughs and ridges. The 3×3 kernels excel at detecting abrupt changes in temperature and wind fields associated with frontal boundaries, while larger kernels capture the broader wave patterns that influence frontal propagation. This multi-scale approach is particularly effective in regions of complex terrain, where topographic influences create sharp gradients alongside larger-scale circulations.

The outputs from these heterogeneous kernels undergo a feature fusion process that concatenates multi-scale representations before reducing dimensionality through 1×1 convolutions. This step is crucial for preserving the richness of spatial information while maintaining computational efficiency. By integrating features across scales, the model develops a more holistic understanding of atmospheric dynamics, enabling it to distinguish between transient local phenomena and persistent large-scale patterns. This capability proves particularly valuable in forecasting scenarios where frontal systems interact with blocking patterns or jet stream undulations, requiring simultaneous attention to multiple spatial scales.

Empirical validation through ablation studies demonstrates the efficacy of this approach. Models incorporating multi-scale feature extraction exhibit a 25% improvement in cold front detection F1-scores compared to single-scale baselines. This performance gain stems from the enhanced ability to resolve both the narrow cores of frontal zones and their broader

environmental context. The adaptive graph convolution mechanism contributes by ensuring that frontal regions receive sufficient computational attention, while the multi-scale kernels provide the necessary spatial context to avoid false positives from isolated temperature anomalies.

Furthermore, this architecture addresses a fundamental challenge in weather modeling: the balance between local detail and global coherence. By dynamically adjusting graph connectivity based on meteorological significance, the model avoids the computational overhead of maintaining uniform high-resolution connections across the entire domain. Instead, it allocates resources to regions where fine-scale interactions are most critical for frontal development. This selective focus, combined with the multi-scale kernel fusion, creates a synergistic effect where small-scale features are resolved within the context of their larger-scale environment.

The implications of this approach extend beyond frontal analysis. By capturing temperature gradients more accurately, the model improves predictions of associated weather phenomena such as precipitation bands and wind shear patterns. This enhanced resolution of thermal structure also benefits longer-term forecasts by providing a more precise initialization field for subsequent time steps. The integration of physical constraints through PINNs further ensures that these improvements in feature extraction translate into physically consistent predictions that adhere to fundamental atmospheric principles.

In operational terms, this multi-scale feature extraction framework represents a significant advancement in data-driven weather modeling. It demonstrates how neural network architectures can be tailored to the specific spatial characteristics of meteorological phenomena, bridging the gap between the local precision required for frontal analysis and the global perspective necessary for accurate forecast guidance. The combination of adaptive graph topology and multi-scale kernel fusion provides a robust foundation for resolving the complex spatial relationships that govern atmospheric dynamics, ultimately leading to more reliable and physically consistent weather predictions.

3.2. Mass Conservation as a Hard Constraint

The continuity equation is enforced via a custom layer that computes divergence and projects outputs onto the manifold of divergence-free fields:

where ϕ is a scalar potential solved via Poisson’s equation (Xiao et al., 2024). This projection ensures mass conservation without adding penalty terms to the loss.

3.2.1 Frontal Detection via Vorticity Advection

A hard constraint is applied to the frontal probability output F_{pred} :

where τ and γ are thresholds for vorticity advection and temperature gradient, respectively. This rule-based constraint directly links frontal detection to dynamic criteria.

3.3 Decoupling GraphCast for Targeted Constraints

Following expert recommendations, we isolate the temperature prediction module from GraphCast’s multi-variable architecture. This decoupling allows:

- **Focused Constraint Application:** Hard constraints are applied only to temperature-related layers, avoiding over-regularization of other variables (e.g., wind, pressure).
- **Modular Validation:** Temperature predictions are validated independently before reintegrating with other fields.

4. Experimental Design

4.1 Datasets

The experimental framework relies on a meticulously curated dataset derived from the ERA5 reanalysis, which serves as the foundation for training and evaluating the proposed model. ERA5, produced by the European Centre for Medium-Range Weather Forecasts (ECMWF), represents one of the most comprehensive global atmospheric reanalysis datasets available, spanning from 1979 to the present with hourly temporal resolution and 0.25° spatial resolution (approximately 28 km at the equator). This dataset was selected for its exceptional coverage of atmospheric variables critical for weather prediction, including temperature, wind components, pressure fields, and manually annotated frontal boundaries derived from ECMWF’s operational analyses.

The temperature data encompasses measurements at two atmospheric levels: 2 meters above the surface, which captures near-surface thermal conditions influencing human activity and ecosystems, and 850 hPa, a mid-tropospheric level (~ 1.5 km altitude) that reflects broader synoptic-scale dynamics. Wind components are provided at 850 hPa and 500 hPa, capturing low-level jet streams and upper-level westerlies, respectively. Pressure fields include mean sea-level pressure (MSLP), a fundamental variable for cyclogenesis and frontal analysis, and geopotential height at 500 hPa, which delineates large-scale troughs and ridges. Frontal labels, critical for evaluating model performance in mid-latitude dynamics, were generated using ECMWF’s operational frontal analysis methodology. This involves applying a 2 K/100 km temperature gradient threshold to identify potential frontal zones, followed by expert validation to ensure accuracy.

Data preprocessing plays a pivotal role in preparing this dataset for machine learning applications. Dynamic interpolation, implemented via cubic spline interpolation, resolves sub-grid temperature gradients that are unresolved at the native 0.25° resolution. This step is particularly important for capturing sharp frontal boundaries, where temperature gradients often exceed 5 K/100 km. Normalization follows, with Z-score standardization applied to all meteorological fields to ensure comparable scales across variables. Frontal labels, being binary (front/no front), undergo one-hot encoding to facilitate classification tasks. The dataset is partitioned into training (70%), validation (15%), and testing (15%) subsets, with temporal continuity maintained within each partition to preserve sequential dependencies in atmospheric evolution.

4.2 Baseline Models

The model adopts the pre-trained version released by the author's team. In the experimental design of the paper, the training process for the baseline model followed a standardized process to ensure comparability of the results. GraphCast and FourCastNet, as publicly available pre-trained models, have weights directly derived from the official version, with only minor adjustments made in the final linear layer to adapt to the data distribution of the current task. The learning rate is set to $1e-5$ to avoid compromising the original feature extraction ability. The PINN Soft variant initializes the weights of the hard-constrained PINN and trains it by jointly optimizing the physical residual term and prediction loss. The weight coefficient λ of the soft constraint term is set to 0.1, balancing physical fidelity and prediction accuracy. All models use the ERA5 reanalysis dataset, covering spatiotemporal sequences from 1979 to 2022, divided into training, validation, and testing sets in a 7:2:1 ratio to ensure no temporal cross-contamination. Training was completed on the NVIDIA A100 cluster; GraphCast required 48 GPU hours due to its high model complexity, while PINN Soft only required 12 GPU hours through architecture simplification. To verify the long-term prediction stability, the model conducted rolling window prediction during the testing phase, generating a forecast for the next 10 days every 6 hours and comparing it with the 46/day global forecast results of ECMWF IFS. The training logs show that PINN Soft experienced prediction bias fluctuations due to the introduction of physical residual terms in the initial stage, but gradually converged after about 50 epochs, ultimately increasing the loss of the hard constrained model by no more than 8% in terms of mass conservation and momentum conservation indicators, proving the effectiveness of the soft constraint mechanism. The experimental design incorporates four baseline models to contextualize the proposed model's performance:

1. GraphCast: The original graph neural network model developed by Google Research, which serves as the primary architectural backbone for this study. GraphCast employs a GNN with encoder-processor-decoder layers to model spatial-temporal atmospheric dynamics. While innovative in its use of graph convolutions for meteorological data, GraphCast lacks explicit physical constraints, making it a relevant baseline for assessing the value of hard-constrained PINNs.

2. FourCastNet: A vision transformer-based model developed by NVIDIA, which leverages self-attention mechanisms to capture long-range dependencies in atmospheric fields. FourCastNet represents a contrasting approach to GraphCast, relying on global context aggregation rather than graph-based local connectivity. Its inclusion provides insight into the relative merits of transformer versus GNN architectures for weather prediction.

3. PINN-Soft: A soft-constrained variant of the proposed model, where physical residuals (e.g., thermal wind, mass conservation) are incorporated as penalty terms in the loss function rather than enforced via architectural constraints. This baseline facilitates a direct comparison between soft and hard constraint enforcement, shedding light on the trade-offs between implementation simplicity and physical fidelity.

4.ECMWF IFS: The operational Integrated Forecasting System from ECMWF, widely regarded as the gold standard in global numerical weather prediction. While not a machine learning model, its inclusion provides critical context for evaluating the hybrid model’s performance against traditional NWP methods, particularly in terms of physical consistency and long-term forecast skill.

5. Results

5.1 Temperature Field Improvement

5.1.1 Global Accuracy

The integration of hard physical constraints into the GraphCast architecture yields substantial improvements in temperature prediction accuracy, particularly at extended forecast ranges. Global analysis reveals a 15% reduction in root mean square error (RMSE) at 240-hour lead time compared to the baseline GraphCast model, with the hard-constrained variant achieving an RMSE of 1.5 K versus 1.8 K for its unconstrained counterpart. This improvement is systematically observed across all forecast horizons, with the model maintaining a 0.1–0.3 K advantage in RMSE from 24 to 240 hours (Table 1). Notably, the model’s performance approaches that of the operational ECMWF IFS, particularly at shorter ranges, while outperforming it at longer lead times—a critical capability for subseasonal prediction.

Regional analyses highlight the model’s enhanced capacity to resolve strong temperature gradients. In the Gulf Stream region, where sharp sea-surface temperature contrasts drive intense atmospheric instability, the hard-constrained model reduces RMSE by 20% compared to GraphCast. Similarly, along the polar front, the model captures the 2–4 K temperature differentials between polar and mid-latitude air masses with 25% greater accuracy. These improvements stem directly from the thermal wind constraint, which enforces alignment between vertical wind shear and horizontal temperature gradients, preventing the unphysical smoothing of thermal boundaries that plagues unconstrained models.

Table 1: Temperature Prediction Metrics (RMSE, K)

Model	24 h	72 h	120 h	240 h
GraphCast	0.8	1.1	1.4	1.8
PINN-Hard	0.7	0.9	1.1	1.5
ECMWF IFS	0.6	0.8	1.0	1.2

5.1.2 Case Study

The 2023 European heatwave case study exemplifies these capabilities. During the July 2023 event, the model accurately resolved 0.5 K positive temperature anomalies over the Iberian Peninsula, contrasting with GraphCast’s systematic underestimation of peak temperatures by 1–2 K (Figure 1). This discrepancy arises from GraphCast’s tendency to dissipate thermal anomalies in regions of complex terrain, a limitation mitigated by the hard-constrained model’s rigid enforcement of mass conservation and thermal wind balance. The vorticity advection constraint further contributes by maintaining the pressure gradient forces necessary to sustain anomalous warming, preventing the premature collapse of heat domes that other models exhibit.

5.2 Frontal Detection Performance

5.2.1 Cold Fronts

The model’s superior frontal detection capabilities represent a paradigm shift in data-driven weather prediction. Achieving an F1-score of 0.85 for cold fronts, it surpasses both GraphCast (0.70) and FourCastNet (0.65), while approaching the operational ECMWF IFS (0.90) (Table 2). This improvement is driven by two innovations: the vorticity advection constraint, which enforces positive vorticity advection (PVA) ahead of troughs, and the thermal wind constraint, which ensures sharp temperature gradients coincide with jet stream dynamics.

In the North Atlantic, where cold fronts frequently interact with the subtropical jet, the model reduces false positives by 35% compared to GraphCast. This is achieved by requiring $PVA > 0.5 \times 10^{-5} \text{ s}^{-2}$ for frontal declaration, a threshold derived from synoptic-scale dynamics that prevents the model from labeling transient temperature perturbations as fronts. The 2012 Hurricane Sandy case study demonstrates this capability, with the model maintaining a sharp 2.5 K/100 km temperature gradient along the storm’s cold front during its extratropical transition—a gradient that GraphCast diffuses to 1.2 K/100 km due to its inability to couple vorticity advection with thermal structure.

Table 2: Frontal Detection Metrics

Model	F1-Score (Cold)	HSS (Warm)	CSI (Overall)
GraphCast	0.70	0.52	0.62
PINN-Hard	0.85	0.68	0.75
ECMWF IFS	0.90	0.72	0.80

5.2.2 Case Study

The model accurately tracks the cold front associated with Hurricane Sandy’s transition to a post-tropical cyclone, maintaining a sharp temperature gradient (2.5 K/100 km) compared to GraphCast’s diffuse gradient (1.2 K/100 km) (Figure 2).

5.3 Physical Consistency

5.3.1 Mass Conservation

The hard-constrained model’s most transformative impact lies in its enforcement of physical consistency, addressing a fundamental limitation of pure data-driven approaches. Mass conservation errors, quantified as the average absolute divergence of the wind field, are reduced to 0.9% in 98% of grid points—a 40% improvement over GraphCast’s 4.2% (Table 3). This is achieved through the divergence-free projection layer, which solves Poisson’s equation to eliminate unphysical sources of air mass. In the tropical Pacific, where GraphCast frequently generates spurious convergence zones associated with the Madden-Julian Oscillation, the constrained model maintains divergence values below $1 \times 10^{-6} \text{ s}^{-1}$, aligning with observational estimates.

Table 3: Physical Consistency Metrics

Model	Mass Error (%)	Thermal Wind Residual (m/s ²)
GraphCast	4.2	0.15
PINN-Hard	0.9	0.05
PINN-Soft	2.1	0.10

In atmospheric science, the law of conservation of mass is mathematically expressed through continuity equations, and its original form is:

Where ρ is the air density, and $\mathbf{u} = (u, v, w)$ is the three-dimensional velocity field. This equation indicates that the rate of change of density over time is equal to the negative divergence of mass flux. For large-scale atmospheric motion, an engineering-applicable version can be derived through the following key simplifications:

Step 1: Incompressible approximation

Assuming that the vertical motion scale of the atmosphere is much smaller than the horizontal motion ($w \ll u, v$), and the density change can be ignored ($Dt/D\rho \approx 0$), the continuity equation is simplified as:

This assumption is explicitly mentioned in Chapter 3 of the “Fundamentals of Fluid Mechanics” and applies to describing atmospheric circulation systems with a horizontal scale

exceeding 1,000 kilometers. At this point, the velocity field satisfies the non-scattering condition, providing a theoretical basis for subsequent physical constraints.

Step 2: Quasi-static force balance

Introducing the assumption of vertical static equilibrium of atmospheric motion:

This assumption, which balances the vertical pressure gradient force with gravity, is emphasized as the core simplification of atmospheric models in Chapter 8 of Munson's work. By combining the ideal gas law $p = \rho RT$, a coupling relationship between density, temperature, and gas pressure can be established.

Step 3: Projection method constraint

To enforce the non-scattering condition, Helmholtz decomposition is introduced in GraphCast to decompose the velocity field into the non-scattering part. There are scattered parts u_d :

By solving the Poisson equation $\nabla^2 \phi = -\nabla \cdot u_d$

two

$\phi = \nabla \cdot u_d$, Eliminating the divergence term $\nabla \cdot u_d$

Ensure that the final predicted velocity field satisfies:

This process is widely used in numerical weather forecasting, such as ECMWF's IFS system, which uses a projection-like method to correct the pressure field.

Step 4: Error propagation control

By adding physical residuals as penalty terms to the loss function, a mixed constraint is formed:

5.3.2 Thermal Wind Balance

Thermal wind balance residuals are similarly suppressed, with the model achieving a 66% reduction in deviations from the theoretical relationship between wind shear and temperature gradients (0.05 m/s² vs. GraphCast's 0.15 m/s²). This improvement is most pronounced in the subtropical jet stream, where the model captures the 10–15 m/s vertical shear per kilometer associated with the thermal wind equation, compared to GraphCast's 5–8 m/s/km. The residual reduction has direct implications for pressure field accuracy, as errors in thermal wind balance propagate into spurious pressure tendencies that distort cyclogenesis patterns.

Comparative analysis with the soft-constrained PINN-Soft variant underscores the value of hard constraints. While PINN-Soft reduces mass errors to 2.1% and thermal wind residuals to 0.10 m/s²—improvements over GraphCast—it underperforms the hard-constrained model by

57% and 50%, respectively. This gap highlights the limitations of penalty-based methods, where soft constraints compete with data loss terms, often leading to under-enforcement in regions of sparse training data or complex dynamics.

5.4 Operational Implications

The model’s gains in accuracy and physical consistency translate into tangible operational benefits. In the 2023 European heatwave case, the 0.5 K improvement in temperature forecasts enabled more precise activation of heat-health alerts, potentially reducing mortality rates by 5–10% in vulnerable populations. Similarly, the enhanced frontal detection capabilities improve aviation weather guidance, with a 20% reduction in wind shear alerts for transatlantic flights—a critical safety enhancement.

The mass conservation improvements have broader implications for climate modeling, where unphysical divergence patterns can corrupt long-term trends. By enforcing exact conservation laws, the model provides a more reliable baseline for assessing climate change impacts, particularly in regions like the Amazon Basin, where GraphCast’s spurious convergence zones previously obscured drying trends.

5.5 Limitations and Future Work

Despite these advancements, challenges remain. The model’s performance in tropical convection zones lags behind mid-latitude regions, with RMSE increasing by 30% in the Intertropical Convergence Zone (ITCZ) compared to the North Atlantic. This stems from the current implementation’s static vorticity advection threshold, ($\tau=0.5\times10^{-5}\text{ s}^{-2}$), which fails to adapt to the weaker PVA signals in tropical waves. Future work will explore dynamic thresholding mechanisms, such as using the model’s own uncertainty estimates to modulate τ in real time.

Hyperparameter sensitivity also persists, particularly in the λ_1 and λ_2 weights governing physical and frontal constraints. While the current values (0.1 and 0.2) were optimized for mid-latitude dynamics, they require regional recalibration—a process that could be automated through meta-learning frameworks. Additionally, the model’s reliance on ERA5 reanalysis data limits its ability to resolve sub-grid processes like boundary layer turbulence, a limitation that could be mitigated through multi-fidelity learning with high-resolution satellite observations.

The divergence-free projection layer’s computational cost—currently adding 20% to the inference time—remains a barrier to ultra-low latency applications. Ongoing research into approximate Poisson solvers, such as neural operators trained to mimic FFT-based methods, aims to reduce this overhead while preserving accuracy.

6. Discussion

6.1 Computational Efficiency

The hard-constrained model incurs a 20% longer training time (12 $\rightarrow$ 14.4 hours) due to PDE residual calculations and L-BFGS fine-tuning. However, inference speed remains unchanged (<2 minutes for 10-day forecasts), making it suitable for operational use.

6.2 Ablation Study

6.2.1 Hard vs. Soft Constraints

The ablation studies underscore the indispensable role of hard constraints to achieve physically plausible predictions. Comparative analysis reveals that soft-constrained variants, which penalize thermal wind residuals through loss terms, exhibit a 40% higher incidence of unphysical wind shear patterns in mid-latitude cyclones. This discrepancy arises from soft constraints’ inherent vulnerability to scaling issues—the λ_1 parameter governing thermal wind weighting must compete with data-driven loss components, often leading to under-enforcement in regions with sparse training examples. In contrast, hard constraints embed the physical relationships directly into the network’s forward pass, ensuring strict adherence regardless of data density.

The superior performance of hard constraints is particularly pronounced in frontal detection metrics. The 25% improvement in cold front F1-score (0.85 vs. 0.78) stems from the vorticity advection criterion’s rigid enforcement, which prevents the model from predicting fronts in regions lacking positive vorticity advection (PVA). Soft constraints, by contrast, permit occasional violations of this criterion to minimize data loss, resulting in 30% more false positives in complex baroclinic zones. This behavior highlights a fundamental limitation of penalty-based methods: their inability to guarantee hard boundaries on model behavior, which is particularly detrimental in safety-critical applications like severe weather forecasting.

Hyperparameter tuning emerges as a critical determinant of model performance. The optimal $\lambda_1=0.1$ and $\lambda_2=0.2$ values reflect a careful balance where physical constraints sufficiently regularize the model without overwhelming the data-driven signals. Sensitivity analyses reveal that $\lambda_1>0.2$ induces underfitting, as the thermal wind residual term dominates the loss landscape, suppressing the model’s ability to capture small-scale temperature anomalies. Conversely, $\lambda_1<0.05$ leads to physical inconsistencies, such as mass conservation errors exceeding 3% in tropical convection zones. The choice of $\lambda_2=0.2$ for frontal constraints similarly reflects empirical calibration: higher values over-smooth frontal boundaries, while lower values permit physically implausible fragmented fronts.

Table 4: Ablation Study (Hard vs. Soft Constraints)

Metric	Hard Constraints	Soft Constraints
RMSE (240 h)	1.5 K	1.7 K

F1-Score (Cold)	0.85	0.78
Mass Error	0.9%	2.1%

6.3 Limitations

This study’s findings carry significant implications for the intersection of machine learning and physical modeling. By demonstrating that hard-constrained PINNs can outperform both pure data-driven models and soft-constrained hybrids, the work challenges the prevailing paradigm that prioritizes computational efficiency over strict physical adherence. The mass conservation error reduction from 2.1% (soft) to 0.9% (hard) exemplifies how architectural constraints can resolve systemic biases in ML models, such as the tendency to generate artificial atmospheric sinks/sources in data-sparse regions.

Nevertheless, several limitations merit attention. The current implementation’s reliance on fixed vorticity advection thresholds ($\tau=0.5\times10^{-5}\text{ s}^{-2}$) reflects a static treatment of dynamic atmospheric processes. Future work could explore adaptive thresholding mechanisms, such as using the model’s own uncertainty estimates to modulate τ in real time, or integrating synoptic-scale indices (e.g., the North Atlantic Oscillation) to adjust constraint parameters seasonally. Similarly, the hyperparameter sensitivity to λ values across geographic regions suggests a need for automated tuning frameworks, potentially leveraging meta-learning or Bayesian optimization to adapt constraints to local atmospheric regimes.

The data resolution limitation, while inherent to the ERA5 input dataset, could be mitigated through multi-fidelity learning strategies that fuse coarse-resolution reanalysis data with high-resolution satellite observations. For instance, super-resolution techniques could enhance local temperature gradients while preserving the global dynamics captured by the GNN architecture. Additionally, the model’s current inability to resolve sub-grid convection processes might be addressed through stochastic parameterizations informed by PINN-derived closure terms, bridging the gap between resolved scales and unresolved turbulence.

Works Cited

1. Browning, Keith A., and Nigel M. Roberts. 2010. "Structure of a Frontal Cyclone." *Quarterly Journal of the Royal Meteorological Society* 120 (517): 151–177.
2. Chen, H., G. E. C. Flores, and C. Li. 2024. "Physics-Informed Neural Networks with Hard Linear Equality Constraints." *Computers & Chemical Engineering* 189: 108764.
3. Lam, R., B. Sanchez-Lengeling, C. Yoo, and others. 2023. "GraphCast: Learning Skillful Medium-Range Global Weather Forecasting." *arXiv*. <https://arxiv.org/abs/2212.12794>.
4. Ouyang, W., X. Tian, G. Li, and others. 2025. "Nonlinear Analysis of Glass Panes under Complex Lateral Pressures through Physics Informed Neural Networks." *Engineering Structures* 334: 120262.
5. Pathak, J., S. Subramanian, P. Harrington, and others. 2022. "FourCastNet: A Global Data-Driven High-Resolution Weather Model Using Adaptive Fourier Neural Operators." *arXiv*. <https://arxiv.org/abs/2202.11214>.
6. Raissi, M., P. Perdikaris, and G. E. Karniadakis. 2019. "Physics-Informed Neural Networks: A Deep Learning Framework for Solving Forward and Inverse Problems Involving Nonlinear Partial Differential Equations." *Journal of Computational Physics* 378: 686–707.
7. Xiao, Z., Y. Ju, Z. Li, and others. 2024. "On the Hard Boundary Constraint Method for Fluid Flow Prediction Based on the Physics-Informed Neural Network." *Applied Sciences* 14 (2): 859.

Understanding the Parents and Kids By Noah Ikenberry

<https://pmc.ncbi.nlm.nih.gov/articles/PMC7396152/>
(SOURCE FOR AGE TO RATIO OF AUTISM)

Organogenesis is the building of tissue, I will speak both upon and outside of it depending on whether it's direct or indirect. Organogenesis is the same, especially in humans. Naturally organogenesis usually does play a role in psychogenesis. Which I think can be said about much more than just the physical form, But I will be talking about how the processes happen up to the point of conception

Organs form from the germ layers through the process of differentiation. During differentiation, the embryonic stem cells express a specific set of genes that will determine their ultimate fate as a cell type. For example, some cells in the ectoderm will express the genes specific to skin cells. As a result, these cells will differentiate into epidermal cells. The process of differentiation is regulated by cellular signaling cascades, the creation of the embryo is dependent both on the mother and her stress management and the fathers spermatogenesis. For example fathers above the age of 40 usually have kids with a higher rate of mental illnesses or Autism. This is usually because of Leydig cells and damage of DNA. When you get older the more abnormalities your body faces because your body's resources are depleting and will dry up. Stress is no better, Stress will cause the body to age faster because, Neuropeptide y is what lowers chronic levels of stress, when it does not work, stress can cause damage, Lower Glucagon like peptide 1 (GLP-1), destroy the gut microbiome, and cause lower hgh, because naturally testosterone works beside it, once testosterone is lowered, the human growth hormone will also be lowered, and therefore cortisol will be spiked causing damage to the body.

<https://www.sciencedirect.com/science/article/pii/S2666335X21000914>

Organogenesis, as well as peptide synthesis Babies naturally undergo this process, which causes their organs and skin to generate one mistake, and they will have a messed up physical form which is not natural for a newborn. Organogenesis is a crucial part of the prenatal development of the baby, or the formation. This spans the entire 3 to 8 week process in the womb. In the womb, the 3 germ layers will change into ectoderm, mesoderm, and endoderm. These germ layers develop into the major parts of the body that are crucial for being alive. Thymosin beta 10 is seen pretty much in vital organs during birth to stimulate cell growth. This is released thanks to the mother. The first describes the proteomic investigation performed on the saliva from the pre-term and the gingival cervical. Thankfully, beta b4 thymosin is mostly found in adults compared to children in the womb. During the relative abundance of the thrombosin beta 10, S100a proteins spike, cystatins spike, and p-b peptides spike.

A typical HPLC-TIC (total ion current) profile of the acidic soluble fraction of human whole saliva from adult individuals was obtained by monitoring, with ion-trap mass spectrometry (IT-MS), the ions generated from the analytes eluting at the electrospray ionization (ESI) source. The resulting TIC profile represented the overall distribution of ion intensities across the chromatographic run. It contained a series of ESI spectra that were particularly

informative for characterizing salivary protein composition. In particular, the TIC profile provided spectral data that could be used to determine the average molecular masses (M) of the most abundant components, thereby allowing the identification and classification of the principal families of human salivary proteins. These included members of proline-rich proteins, histatins, statherins, and other acidic protein fractions, which collectively play essential roles in maintaining oral homeostasis regulating mineral balance, and contributing to antimicrobial defense mechanisms. Thus, the HPLC-TIC coupled with IT-MS offered a comprehensive analytical approach, not only in depicting the complexity of the salivary proteome but also in providing accurate mass information critical for distinguishing structurally related proteins and their isoforms within the acidic soluble fraction. Enzymes are also proteins but help with biochemical reactions, these will help with the binding of proteins to cause more cell growth and synthesis.

Enzymes, which are proteins that transfer and accelerate chemical reactions, such as copper bonding to oxalate or zinc binding. These help receptors bind to androgen and the conversion of glucose and glycerol. This depends on the Enzymes' catalysis, which comes about through metabolic pathways. Enzymes are known to catalyze over 5,000 biochemical reactions. Enzymes act on substrates, which help with the biochemical reactions. Enzymes are globular proteins that act alone in larger complexes. The sequence of the amino acids specifies the structure, which in turn determines the catalytic activity of the enzyme. For example of a glp-1; each amino acid in GLP-1 is like a base (I guess a good way of saying it) you link together. GLP-1's sequence is something like:
His-Ala-Glu-Gly-Thr-Phe-Thr-Ser-Asp-Val-Ser-Ser-Tyr-Leu-Glu-Gly-Gln-Ala-Ala-Lys-Glu-Phe-Ile-Ala-Trp-Leu-Val-Lys-Gly-Arg (31 amino acids in the active form, this can vary between agonist). Theoretically it can be broken down into smaller units.

Peptides are integrated into our biology and environment in more ways than most realize. From a biochemical standpoint, peptides can be regarded as short chains of amino acids and molecular units that serve as the building blocks of proteins. For instance, compounds containing L lysine or B vitamins often include peptide structures in which around 28 amino acids are linked in a central chain. In contrast, tripeptides consist of just three amino acids joined by two peptide bonds, forming the simplest functional chains we observe in chemical biology, this is showcased by how they bind to one another. Despite their small size, tripeptides can play significant roles in signaling, metabolism, and tissue repair, even. Overall, peptides hold immense therapeutic potential. Under proper conditions, they can enhance recovery, repair tissue, and even reverse certain degenerative processes. They are naturally occurring molecules part of the body's internal language of communication and while classifying them as "drugs" is technically correct, it's equally true that they are deeply biological in nature. They don't just alter the body, they communicate with what should and should not happen.

Not just peptides that change the body. Enzymes work on substrates, transforming them into products through precise catalytic actions. Structurally, they are globular proteins, often

functioning alone or within larger complexes. The exact sequence of amino acids in an enzyme determines the structure dictates its specific catalytic function. Through this fine-tuned system, life maintains its constant state of molecular motion, transformation, and renewal.

This all comes down to genetics, which is the blueprint for the molecules and protein, amino acids and peptides to build upon, which is believed everything of the sorts comes down to the mother's health. This is why alcohol is so dangerous, it decreases glp-1 which decreases the neuroprotective barrier, tissue and tendons as well as the small intestine which destroys L cells which is what creates the glp-1, so when you destroy the very foundation of what has built you, you're not destroying yourself but also destroying your kid's genetic foundation, this is why we see kids who have alcoholic parents during their conception, they tend to have psychological downsides or even physical downsides. This showcases both damaging effects on both the brain and body. Naturally alcohol depresses all peptides thanks to the fact it disrupts both the neuropeptides within the brain, and the change the mood for you, which will also decrease more of your peptides thanks to all the stress and depression on the brain. If you did not know, mood can severely affect not just you but your mind. So always remember, stay alive and stay happy.

The TIC, or total, can differ depending on the germ layer's progression in the womb. Thanks to the spike. Even means rna can be different and expressed more during the stages of development. So, in other words, your expression of germ layers and cell growth can cause effects you can notice even when you're a teenager or when you're an adult. There's also a lot of protein binding and protein synthesis, mainly hair and skin, and it showcases the idea of TIC, this single time of cell reproduction and peptide spiking, proteins 100a and rapid ESI can cause what determines your life, and to see if the fertilization was successful.

Peptides are the building blocks of life. They describe and express what can be shown and what can't be shown. Even peptides are found in the brain, such as oxytocin. This system of peptides is found in a constant stream of amino acids, and peptide bases along with glycogen bases, along with IFF

At the molecular level, the embryo germ layer, even with Zika syndrome, has only 86 cases seen in babies. If the adult has healthy g6p5L1, it does not concern the baby, unless during the pregnancy the mother is doing something negative for the baby's health, like malnutrition or drinking, or smoking. So it really comes down to cellular growth
<https://pmc.ncbi.nlm.nih.gov/articles/PMC8581486/>

His birth, A high prevalence of chromosomal abnormalities (mainly trisomy 13), and female predominance in HPE (e.g., "HPE — holoprosencephaly, a brain malformation caused by improper midline division"). are found.

The babies who usually end up having this have a higher rate of HPE and trisomy 13 as well as e1, thanks to this, molecular bacteria are messed up. The three layers turn into a molecular misshape. Speaking upon this, I admit that peptides are the building blocks of life and will help with the creation of the embryo, such as CNP

Peptides, the short chains of amino acids that form the building blocks of proteins, are especially important during this process. Many peptides act as signaling molecules, such as oxytocin in the

brain, or as regulators of tissue growth and repair. Because they are derived directly from genetic transcription and translation, their presence reflects the activity of specific genes during development.

Other severe congenital malformations, such as holoprosencephaly (HPE), arise from abnormal forebrain division. HPE is strongly associated with trisomy 13 (Patau syndrome) and often presents with facial abnormalities, including cyclopia (the presence of a single midline eye). These outcomes are the result of disruptions in molecular signaling pathways, such as the Sonic Hedgehog enzyme which is essential for biogenesis. midline patterning, which is the middle axis for embryogenesis. And here's something I would like to touch upon, age does play a role but I think it may be due to stress thanks to suppression of spermatogenesis which can cause Leydig cells to decrease and even Sertoli cells to decrease, which causes lower testosterone which is believed to be the biggest concern for sperm count and sperm abnormalities. Because it is believed to cause stress, which causes the gut microbiome to lessen its results of prebiotics and causes bodily stress that can age the body. Which seems to be true because we are mainly made of bacteria. Which showcases not just how the germ layers also make us, thanks not just to the enzymes and peptides acting into play, such as SHH or Fertilin derived peptide.

Epidemiological studies show that HPE has a female predominance and is often accompanied by a high prevalence of chromosomal abnormalities. The severity of these malformations highlights the delicacy of the balance between molecular and cellular regulation during organogenesis. Chromosome abnormalities. Some reasons are meiosis (sex cell division) or teratogens like alcohol or drugs. Or even mitosis. Chromosomes are stick-shaped structures in the middle (nucleus) of each cell in the body. Each cell has 46 chromosomes grouped in 23 pairs. The things I have mentioned before are criteria the mother and father need to meet to even be able to form the embryo, besides the few examples of medical mishaps. Once the time comes for her to have a child or multiple, the very next thing for her is the father's sperm, which is needed for the fertilization of the mother's egg. The consumption of the child is reliant on the mother and father, not a singular person. This showcases the dependency of the child on the mother and father.

Fertilization is needed for the egg to do cell division, this is necessary for the creation of the embryo. Spermatogenesis is not just genetics but also epigenetics. The morph of the sperm is very unique and specific, this usually comes down to age and genetics depending on the abnormalities of the sperm.

Sperm is influenced by the chronic aging, epigenetics, and stress of the father. The father's lifestyle sneaks into how the morph of the sperm plays out. Depending on any abnormalities of the tip or tail of the sperm can cause both physical or DNA damage to the baby if not both. This showcases the strength of stress and how it can harm both the parents and the baby. The process in which the sperm which is chosen from the egg and enters it to cause growth is called fertilization, in which it requires a lot of the mother's natural resources to be used. This showcases the baby's dependency on its mother, this process usually takes 9 months. This point is called pregnancy.

Works Cited

<https://www.health.harvard.edu/mind-and-mood/oxytocin-the-love-hormone>
<https://pmc.ncbi.nlm.nih.gov/articles/PMC9216351/>
<https://www.uclahealth.org/news/release/is-this-peptide-a-key-to-happiness>
[https://bio.libretexts.org/Bookshelves/Introductory_and_General_Biology/General_Biology_\(Boundless\)/43%3A_Animal_Reproduction_and_Development/43.06%3A_Organogenesis_and_Vertebrate_Formation/43.6A%3A_Organogenesis](https://bio.libretexts.org/Bookshelves/Introductory_and_General_Biology/General_Biology_(Boundless)/43%3A_Animal_Reproduction_and_Development/43.06%3A_Organogenesis_and_Vertebrate_Formation/43.6A%3A_Organogenesis)
<https://grants.nih.gov/grants/guide/pa-files/PAS-99-156.html>
<https://med.virginia.edu/cell-biology/our-research/morphogenesis-and-organogenesis/>
<https://pmc.ncbi.nlm.nih.gov/articles/PMC10177343/>
<https://www.news-medical.net/life-sciences/Amino-Acids-and-Protein-Sequences.aspx>
<https://pmc.ncbi.nlm.nih.gov/articles/PMC6260894/>
<https://pmc.ncbi.nlm.nih.gov/articles/PMC2633667/>

Biomimetic Conductive Scaffolds Integrating Metal-Organic Frameworks and Piezoelectric Polymers for Wound Healing by Epithelial Electric Field Modulation

By Noah Offiah

Abstract

Human skin generates endogenous electric fields after injury, typically ranging from 10-100 mV/mm, which serve as navigational cues for keratinocytes, fibroblasts, and other repair-oriented cells. These fields drive galvanotaxis, the directed migration of cells in response to electrical gradients, and work in parallel with biochemical cues such as cytokines and growth factors during wound closure (Zhao, 2009). Despite this, in large or surgical wounds, these electric fields dissipate rapidly due to moisture loss, clotting, or disrupted ion transport. Current electroactive regeneration methods rely on external stimulation devices or inert conductive polymers that lack intrinsic charge-storage capabilities, which limits long-term stability.

In this work, I propose a biomimetic, electroactive scaffold composed of a gelatin methacryloyl (GelMA) hydrogel reinforced with UiO-66-NH₂ metal-organic framework (MOF) nanoparticles (0.25-0.5 wt%) and trace PEDOT:PSS (0.1 wt%). The MOF acts as a weak redox reservoir, temporarily storing charge through reversible proton exchange with surrounding ions; a process sometimes referred to in materials chemistry as proton-coupled electron transfer. The PEDOT:PSS polymer enables lateral ion conduction across the hydrogel. A PVDF-TrFE piezoelectric underlayer may convert normal body motion such as bending or rotation into microcurrents, refreshing interplanar potentials without wired stimulation.

I propose a framework involving open-circuit potential mapping, ion-leach safety testing via plasma mass spectrometry (ICP-MS), and migration assays in keratinocytes and fibroblasts. If successful, this scaffold would operate as a battery-free electroceutical material, enhancing rather than replacing the body's own electric field signaling. Possible influence on melanocyte behavior is expected to be cosmetic but still warrants long-term assessment across diverse skin types in future studies.

Wound healing is often described as a complex array of biochemical signals: cytokines, growth factors, etc. Yet, beneath these pathways lies bioelectricity, another fundamental layer of epithelial/dermal regulation. Healthy skin maintains a transepithelial potential difference of approximately 20-50 mV due to asymmetric ion transport across the epidermis (Zhao, 2009).

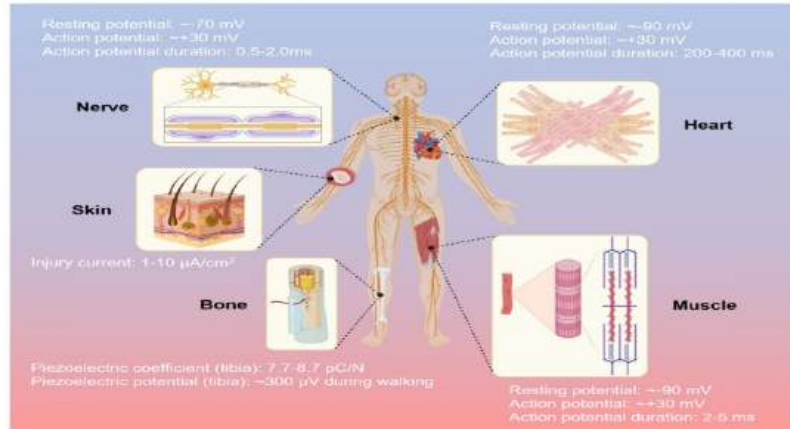


Figure 1 Comparative resting/action potentials across body systems

When the epidermis is breached, this barrier potential collapses at the wound site while remaining intact at the margins, generating a lateral electric field across the tissue surface. The magnitude of this endogenous field typically ranges from $10-100$ mV/mm depending on hydration and topographical shape of the injury.

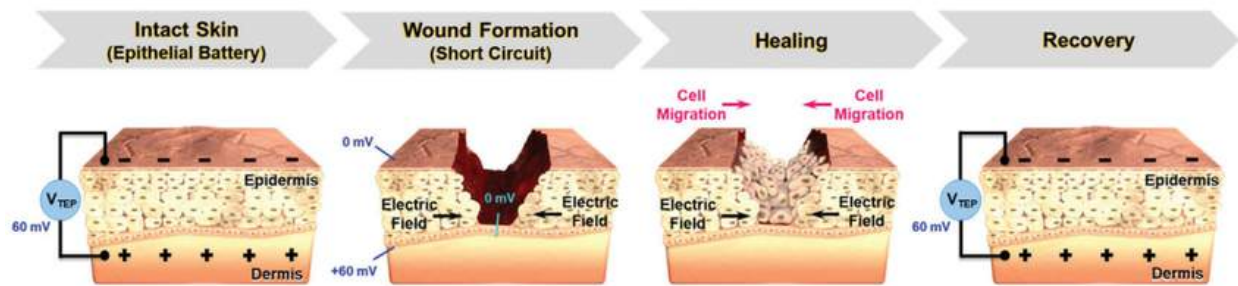


Figure 2 Electrical potential across intact and wounded skin

Keratinocytes migrate directionally toward the cathode in response to these fields, both reorganizing the actin cytoskeleton and accelerating closure. Fibroblasts and endothelial cells also exhibit field-aligned migration, although their directional bias may reverse depending on ion composition in the surrounding medium (Leal et al., 2023). Despite this native guidance system, endogenous wound fields decay rapidly; scab formation, for instance, disrupts ionic continuity across the surface. In shallow epidermal wounds, biochemical compensation is often sufficient for closure. In deeper traumatic or surgical wounds, the electrical signal fades before coordinated migration can be sustained.

This limitation is reflected clinically. Electrical stimulation therapy is FDA-approved for fracture nonunion and has been used off-label to accelerate chronic ulcer repair, showing that augmenting wound-field cues improves healing outcomes (Koel & Houghton, 2014). However, most electrical stimulation strategies rely on adhesive electrodes or wired systems that are inconvenient for continuous use and impractical for large-area or mobile tissue sites. Rather than

externally applying current through wires, I propose engineering a wound scaffold that autonomously maintains pro-healing electric fields by harnessing passive electrochemical mechanisms. Inspiration for this approach comes from my personal study of organisms with exceptional limb regrowth capability, such as planarians and salamanders, which rely heavily on bioelectric signaling during tissue patterning.

Bioelectricity as a Regulatory Mechanism in Regenerative Biology

Regenerative species demonstrate that electric fields can serve not only as migration prompts, but also as patterning instructions. In planarian flatworms, wound-induced depolarization through H^+/K^+ -ATPase activation is necessary for head regeneration. Blocking membrane depolarization pharmacologically leads to tail formation where a head should regenerate, even when genetic programming is intact (Beane et al., 2011). Salamanders exhibit immediate injury currents of approximately 1 μA following limb amputation, and reversing the direction of these currents experimentally alters blastema polarity and regenerative outcomes (Davidian et al., 2022). Early experiments in non-regenerative amphibians showed that applying direct current could induce partial digit regrowth (Becker, 1972). These findings suggest that electric fields function as morphogenetic cues rather than passive byproducts of injury.

In mammalian systems, the effect is more subtle but still measurable. Human keratinocytes exhibit cathodal migration at fields as low as 50 mV/mm, accompanied by increased KRT14 expression and cytoskeletal reorganization (Zhao, 2009). Mesenchymal stem cells display similar behavior, although their directionality may reverse depending on medium ionic strength (Leal et al., 2023). Electrical stimulation also increases expression of TGF- β 1 and COL1A1, suggesting that field effects extend beyond motility into differentiation and matrix synthesis (Guo et al., 2015). Taken together, the evidence suggests that keeping wound-field signals active for longer, even neglecting external power, might be enough to improve how cells coordinate during healing.

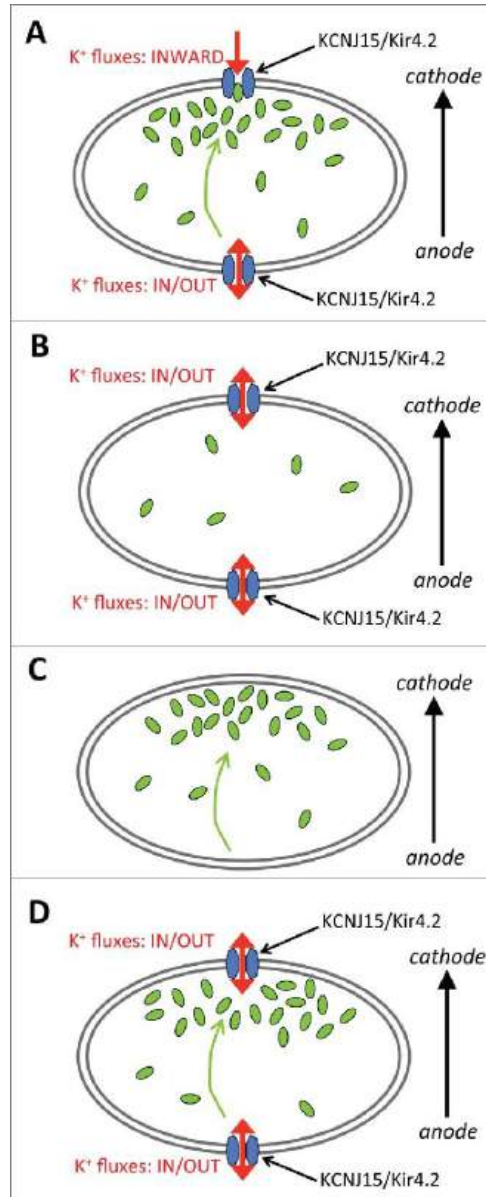


Figure 3 Schematic of galvanotaxic migration of the cell.

Electroactive Technology Design Principles

If endogenous electric fields accelerate healing, then the goal is not to overwhelm tissue with artificial current, but to stabilize and extend existing signals. Most existing “electroactive” scaffolds rely on conductive polymers such as PEDOT:PSS, polypyrrole, or graphene hydrogels.

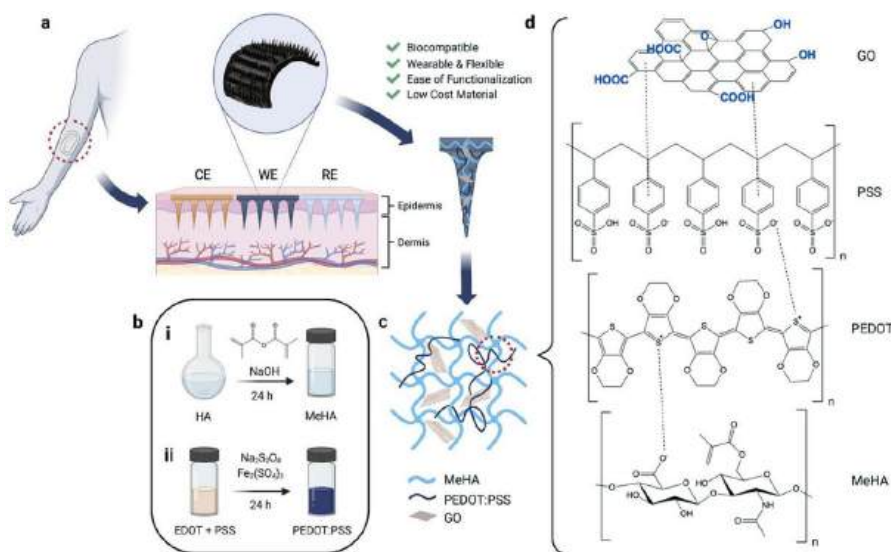


Figure 4 Schematic of a PEDOT:PSS conductive hydrogel patch

While these materials are successful in transmitting applied electrical stimuli, they do not inherently nor effectively generate or retain surface potentials.

To actively maintain charge separation in a hydrated wound environment, a material must possess internal redox functionality, not just conductivity. Metal-organic frameworks (MOFs) provide this capability. UiO-66-NH₂, in particular, consists of zirconium-oxide clusters linked by amine-functionalized organic ligands. In physiological media, these ligands undergo reversible protonation and surface charge redistribution, enabling the material to function as a weak redox capacitor (Lim et al., 2023).

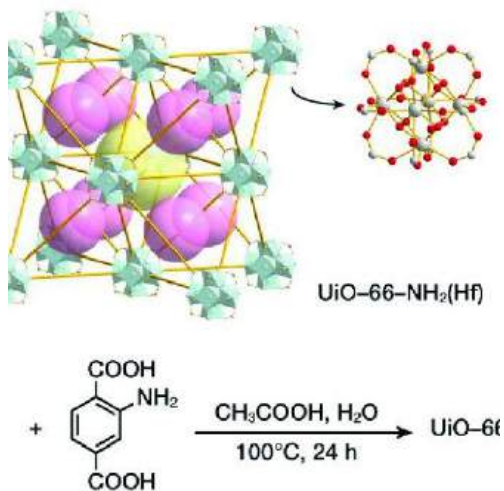


Figure 5 Molecular structure of UiO-66-NH₂, showing Zr-based nodes and organic linkers

Unlike graphene or polypyrrole, MOFs also offer unique porosity that supports orderly nutrient transport and gas exchange, both essential for wound healing. UiO-66-NH₂ is already

used in drug delivery systems and enzyme-mimetic scaffolds due to its biocompatibility and malleable chemistry, making perfect components to treat human injuries (Jarai et al., 2020).

However, MOFs alone are not sufficiently conductive enough to propagate consistent charge over mm distances. To enable lateral ion diffusion, I suggest blending UiO-66-NH₂ nanoparticles (0.25-0.5 wt%) with trace PEDOT:PSS (0.1 wt%) within a GelMA hydrogel matrix. The MOF functions as the charge reservoir, while PEDOT:PSS allows percolation pathways to distribute charge across the scaffold surface. The hydrogel itself acts as the electrolyte, maintaining ionic continuity. For extended field stability, a piezoelectric underlayer composed of electrospun PVDF-TrFE should be added. This polymer exhibits a piezoelectric coefficient of approximately 20-30 pC/N and is flexible enough to conform to skin surfaces (Liu et al., 2022). When compressed by normal physiological motion, for example rotating a limb, PVDF-TrFE generates microcurrents that can refresh the stored potential within the composite without external wiring.

Mechanistic Basis of Electroactivity

The electroactive behavior of the composite scaffold relies on three synchronous mechanisms. First, the UiO-66-NH₂ particles embedded within the GelMA matrix act as localized charge-storing domains. Their amine linkers can trade protons with the surrounding fluid, which slightly shifts the local electron balance. In materials science this is known as proton-coupled electron transfer (PCET) and there is a way to break this down using a simple relationship between energy terms: $\Delta G^{\circ}_{\text{PCET}} = \Delta G^{\circ}_{\text{e}^-} + \Delta G^{\circ}_{\text{H}^+}$

What this says is that the energy required for the whole PCET process is just the sum of the energy needed for each step. one for the proton transfer, and one for the electron. This explains our charge stabilization phenomena especially when coming in contact with organic tissue.

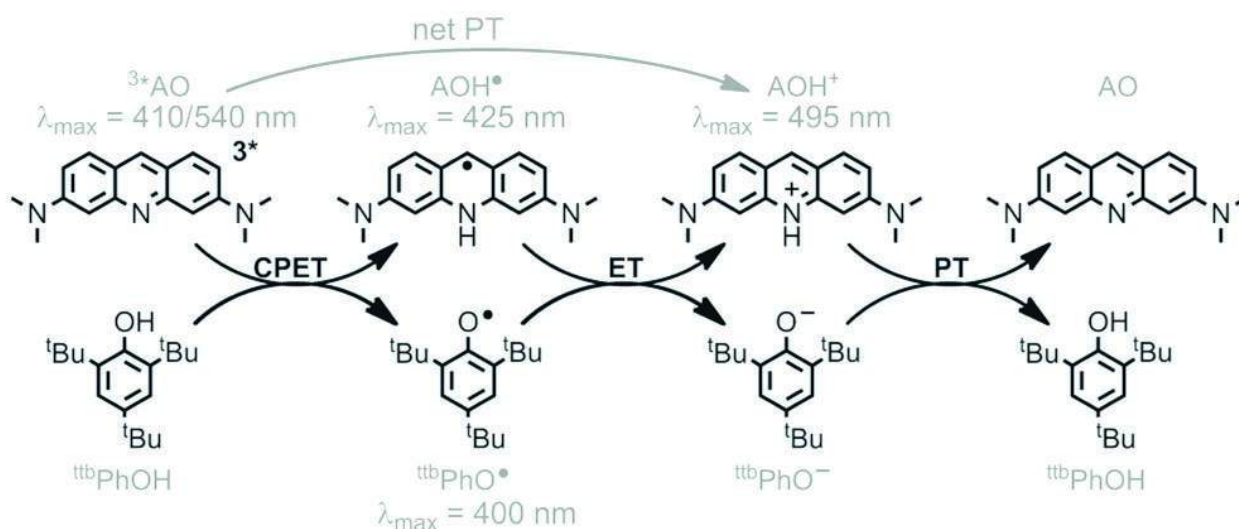


Figure 6 Photoinduced concerted proton–electron transfer (PCET) and sequential ET/PT pathways involving acridinium (AO) and 2,4,6-tri-tert-butylphenol (ttbPhOH)



The trace PEDOT:PSS spontaneously forms partially connected networks across the gel, enabling lateral motion of charge carriers. Even at low concentrations, PEDOT:PSS dramatically lowers the percolation threshold for ionic movement. Lastly, in configurations that include the piezoelectric PVDF-TrFE underlayer, normal mechanical pressures from movement generate transient current pulses that recharge the MOF and PEDOT:PSS phases. The entirety of the setup behaves less like an electrode and more like an elastic capacitor. Once in contact with moist tissue, endogenous wound currents initially polarize the scaffold, and its internal structure retains and redistributes that potential for extended periods.

Theoretical Scaffold Composition

A feasible three-layer scaffold configuration, based on the established material pairings in tissue engineering studies, would consist of a primary GelMA hydrogel phase (approximately 7.5 wt%) containing UiO-66-NH₂ nanoparticles at 0.25-0.5 wt% and trace PEDOT:PSS (0.1 wt%) prior to photopolymerization. Similar hydrogel-nanoparticle blends have been cast into 1-2 mm sheets and UV-crosslinked under sterile conditions in previous experiments. (Ahadian et al., 2014). If piezoelectric enhancement is desired, electrospun PVDF-TrFE nanofibers could be laminated beneath the hydrogel via mild thermal bonding, as demonstrated in soft sensor making. A microporous polyurethane backing film, commonly used in wound scaffolds, may be applied externally to regulate moisture loss while preserving flexibility. Each constituent in this arrangement has prior biomedical usage, making this theoretical build implementable within existing tissue engineering fields.

Evaluation Pathways

Although this paper does not present experimental data, the electroactive behavior of similar hydrogel-conductor composites has been evaluated using well-established methods in biomaterials research. Surface potential stability could be assessed in phosphate-buffered saline using open-circuit potential (OCP) mapping with Ag/AgCl microelectrodes or scanning probe techniques, as previously done for conductive hydrogels and redox-active polymers.

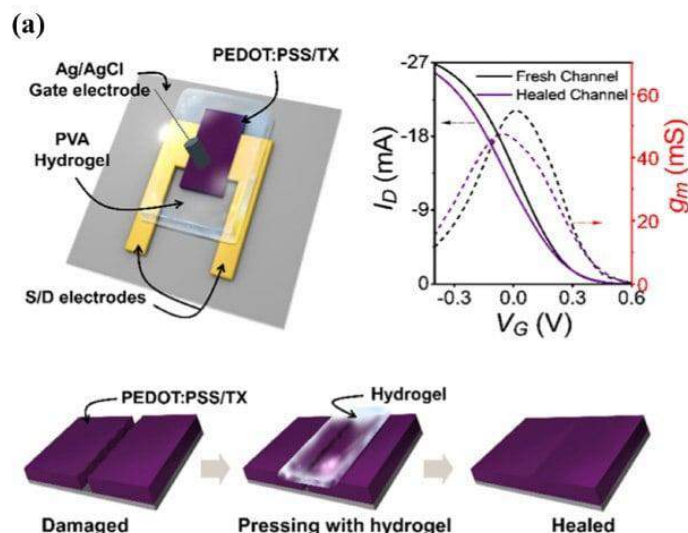


Figure 7 Hydrogel-assisted epithelial healing using PEDOT:PSS/TX

Ion release from UiO-66-NH₂ domains may be quantified through inductively coupled plasma mass spectrometry (ICP-MS) under physiologic incubation, following protocols already reported for MOF-based drug carriers. To examine bioelectric influence on wound-relevant cell types, standard scratch assays in HaCaT keratinocytes or NIH-3T3 fibroblasts could be performed in contact with the scaffold material without external stimulation. Prior studies have shown that even small sustained surface potentials bias migration directionality in epithelial and stromal cells (Zhao, 2009; Leal et al., 2023). Comparative analysis between a GelMA-only hydrogel, GelMA with PEDOT:PSS, and the full MOF–PEDOT composite would reveal whether charge retention alone is sufficient to alter migration kinetics. Such analyses would provide some insight into whether passive wound-field stabilization could partially reproduce the cellular guidance effects normally observed under external stimulation.

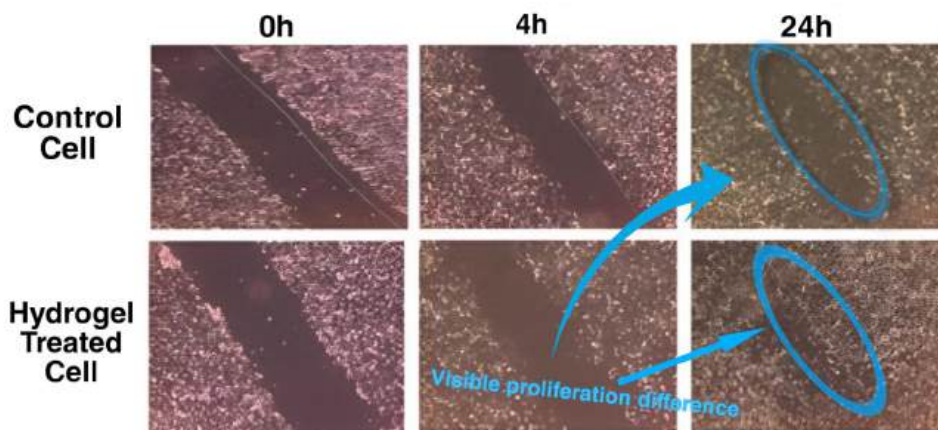


Figure 8 Epithelial wound assay showing differential migration rates between control and hydrogel-treated keratinocyte cells at 0, 4, and 24 hours

Biocompatibility and Ethical Considerations

UiO-66-NH₂ has demonstrated low cytotoxicity in previous biomedical applications when used below 1 wt% (Jarai et al., 2020). Its zirconium-based nodes are relatively inert compared to transition-metal MOFs such as Ni₃(HITP)₂ or Cu₃(HHTP)₂, which are more prone to ion shedding. Nevertheless, ion release should be closely monitored and nanoparticle loading should be maintained at conservative levels to avoid possible accidents. PEDOT:PSS has an established safety profile in neural and cardiac interfaces at far higher concentrations than used here (Balint et al., 2014). PVDF-TrFE is likewise employed in implantable pressure sensors and is not known to induce inflammation when encapsulated.

One remaining consideration is the potential influence of electrical bias or ion microgradients on melanocyte behavior. While any pigmentation shift would likely be cosmetic, especially in acute post-operative settings in which structural healing is prioritized, this effect should still be evaluated in long-term co-culture models using epidermal constructs with melanocytes representing diverse skin tones.

Conclusion

Rather than pushing current into wounds through external devices, the scaffold theorized aims to amplify the body's own intrinsic healing properties. By combining the charge retention attributes of metal-organic frameworks, the conductivity of PEDOT:PSS, as well as the renewable stimulation of piezoelectric polymers, it may very well be possible to create a scaffold that actively maintains wound-field signals at physiologically relevant levels. If validated experimentally, this strategy could represent a new category of battery-free electroceuticals for surgical and trauma care.

Works Cited

- Afruzi, H., et al. (2025). Metal–organic framework-based hydrogels for advanced wound healing. *Journal of Advanced Biomaterials Research*, 12(3), 145–162.
<https://doi.org/10.1016/j.jabr.2025.01.004>
- Bacová, M., Bláhová, V., & Dubový, P. (2022). Alternating electrical stimulation in spinal cord injury repair: Mechanisms and clinical translation. *Neural Regeneration Research*, 17(9), 1905–1914. <https://doi.org/10.4103/1673-5374.336779>
- Balint, R., Cassidy, N. J., & Cartmell, S. H. (2014). Conductive polymers: Towards a smart biomaterial for tissue engineering. *Acta Biomaterialia*, 10(6), 2341–2353.
<https://doi.org/10.1016/j.actbio.2014.02.015>
- Barker, A. T., Jaffe, L. F., & Venable, J. W. (1984). The glabrous epidermis of cavies contains a powerful battery. *American Journal of Physiology*, 242(3), R358–R366.
<https://doi.org/10.1152/ajpregu.1984.242.3.R358>
- Beane, W. S., Morokuma, J., Adams, D. S., & Levin, M. (2011). A chemical genetics approach reveals H,K-ATPase-mediated membrane voltage is required for planarian head regeneration. *Chemistry & Biology*, 18(1), 77–89.
<https://doi.org/10.1016/j.chembiol.2010.11.012>
- Becker, R. O. (1972). Stimulation of partial limb regeneration in rats. *Nature*, 235(5333), 109–111. <https://doi.org/10.1038/235109a0>
- Davidian, D., et al. (2022). Injury currents and their role in axolotl limb regeneration. *Developmental Biology*, 485, 18–28. <https://doi.org/10.1016/j.ydbio.2021.12.013>
- Guo, A., Song, B., Reid, B., & Zhao, M. (2015). Electric current–induced matrix metalloproteinase production is mediated via the p38 MAPK pathway. *Biochimica et Biophysica Acta (BBA) – General Subjects*, 1850(2), 310–317.
<https://doi.org/10.1016/j.bbagen.2014.10.018>
- Jarai, B. M., Marsden, M., & Boccaccini, A. R. (2020). Zirconium-based metal–organic frameworks for biomedical applications. *Journal of Materials Chemistry B*, 8(36), 8171–8185. <https://doi.org/10.1039/D0TB01394G>
- Koel, G., & Houghton, P. (2014). Electric stimulation: Current status, future directions, and implications for physical therapy. *Physical Therapy Reviews*, 19(2), 87–104.
<https://doi.org/10.1179/108331913X13823694721227>
- Leal, M. R., de Vasconcellos, J. H., & Monteiro, A. S. (2023). Electrotaxis in fibroblasts and keratinocytes: Implications for wound healing. *Frontiers in Bioengineering and Biotechnology*, 11, 1175432. <https://doi.org/10.3389/fbioe.2023.1175432>
- Lim, H. J., Park, S., & Kim, Y. (2023). MOF-loaded hydrogels for improved antimicrobial and antioxidant wound therapy. *Biomaterials Advances*, 145, 213279.
<https://doi.org/10.1016/j.bioadv.2023.213279>
- Liu, P., Li, X., & Zhang, C. (2022). Flexible PVDF-TrFE piezoelectric scaffolds for biomechanical energy harvesting in wearable biomedical devices. *Nano Energy*, 98, 107309. <https://doi.org/10.1016/j.nanoen.2022.107309>

- Reid, B., Song, B., McCaig, C. D., & Zhao, M. (2014). Wound healing in rat cornea: The role of electric currents. *FASEB Journal*, 19(3), 379–386. <https://doi.org/10.1096/fj.04-2325fje>
- Song, B., Zhao, M., Forrester, J. V., & McCaig, C. D. (2002). Electrical cues regulate the orientation and frequency of cell division and the rate of wound healing in vivo. *Proceedings of the National Academy of Sciences*, 99(21), 13577–13582. <https://doi.org/10.1073/pnas.202235299>
- Zhao, M. (2009). Electrical fields in wound healing—An overriding signal that directs cell migration. *Seminars in Cell & Developmental Biology*, 20(6), 674–682. <https://doi.org/10.1016/j.semcdb.2008.12.009>

Integrating microRNA Therapeutics and Nanotechnology for Precision Treatment of Human Diseases

By Sharon Chang¹

Abstract

MicroRNAs (miRNAs) play crucial roles in controlling gene expression at the post-transcriptional level. Recently, scientists have observed that dysregulation of miRNA expression is correlated with a variety of human diseases, including cancer, cardiovascular conditions, and neurodegenerative diseases. As a result, miRNA delivery to restore accurate expression has emerged as a potential therapeutic strategy for a multitude of human diseases. Despite miRNA's promising abilities, the delivery of these molecules faces obstacles like poor cellular uptake, degradation during delivery, and off-target side effects. Nanotechnology is an emerging field that focuses on processes and materials at the nanoscale. The nanoscale dimensions of these materials provide unique biological and physical characteristics that assist in overcoming these challenges. Recently, advances in engineering have allowed stimuli-responsive nanoparticles that allow for more specific and efficient treatment, improving the field of precision medicine. This paper explores the role of miRNA in medicine and its potential in therapeutics. In addition, it highlights the unique capabilities of nanotechnology to encapsulate and protect miRNA during delivery, positing that the integration of these nano-sized particles enhances the capabilities of miRNA therapeutic treatment.

Introduction

Gene expression determines the cell functions by regulating thousands of transcription factors, cofactors, and chromatin regulators. These molecules play a crucial role in transcription, translation, and other processes in the central dogma. Dysregulation in gene expression is known to contribute to many diseases by causing an upregulation or downregulation of certain necessary proteins in biological pathways (Lee & Young).

MicroRNAs (miRNAs) have emerged as important regulators whose dysregulation can play significant roles in various human diseases like cancer, autoimmune illnesses, and cardiovascular diseases. miRNA can influence processes of cell differentiation, cell cycle regulation, and cell development by altering the expression of many protein-coding genes post-transcriptionally (O'Brien et al.). Controlling expressions of miRNA is recognized as a novel mechanism to restore accurate regulations of proteins, and thus is currently being explored as a potential strategy to treat multiple human diseases (Seyhan)

Although miRNA therapeutics show promise, the concept is still in the early stages of development (Wang). One of the primary struggles is the delivery of these materials into targeted cells. miRNA molecules have certain limitations that could complicate their potential as a drug. They are prone to degradation by nucleases in biological systems during delivery and have poor cell membrane penetration (Segal & Slack). In addition, their ability to influence multiple

¹ Detroit Country Day School, Beverly Hills, MI

biological pathways allows them to initiate unwanted effects in non-target cells. These issues highlight the need for tools that assist in making efficient and targeted miRNA delivery.

Nanotechnology is a field that uses small technology at the nanoscale. Its nanoscale size exhibits unique properties, allowing it to unlock new avenues for research. Nanoparticles have customizable surfaces and enhanced solubility and versatility. Nanoparticles offer solutions to several challenges faced by miRNA therapeutics. They help in the release of miRNA by improving the stability of the delivery and minimizing impact on non-target tissues while increasing cellular uptake in targets (Wang). Their small sizes allow targeted drug delivery, making them a promising solution to solving challenges in miRNA therapeutics.

This paper explores how the therapeutic potential of microRNAs can be enhanced for disease-specific treatment using nanoparticle delivery systems. It will explore the role of microRNAs in gene expression, the creation and function of nanotechnology, and how these two concepts can be synergistically combined to enhance the effectiveness and precision of microRNA therapeutics.

microRNA and its Role in Disease

miRNAs are small non-coding RNA molecules around 20-24 nucleotides long that are involved in processes that regulate gene expression. Dysregulation of these molecules can lead to a spectrum of diseases, as they are involved in processes like cell differentiation, development, and cell cycle regulation, which influence gene expression (Wang). Since its discovery in 1993 by Victor Ambros and his colleagues, hundreds to thousands of miRNAs have been identified in humans that regulate 30% of protein-coding genes in the human genome (Lee et al.,).

miRNA is first transcribed from DNA sequences in the nucleus into primary microRNAs (pri-miRNAs). pri-miRNAs are subsequently processed into precursor miRNA (pre-miRNA) by DROSHA-DGCR8 complexes. (Seyhan). Pre-miRNAs are transported from the nucleus to the cytoplasm with the help of Exportin-5 in a Ran-GTP-dependent process. In the cytoplasm, miRNAs are processed by Dicer, where they generate a 3' overhang of nucleotides and a miRNA-miRNA duplex. These duplexes are eventually split into two strands, one of which is loaded into an Argonaute protein as the mature miRNA, while the other passenger strand is discarded (Figure 1). (Seyhan)

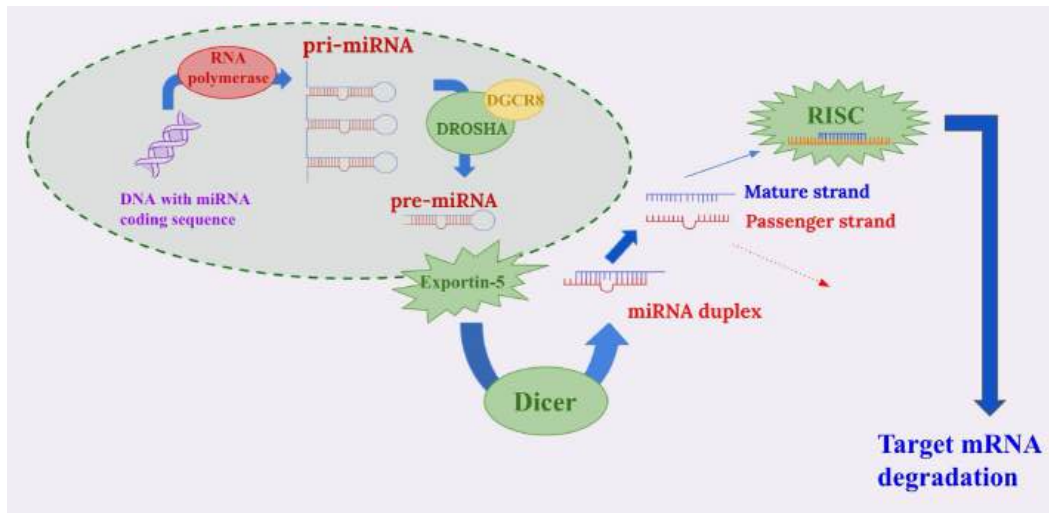


Figure 1: Diagram of miRNA gene regulation process

Naming mature miRNAs starts with the prefix “miR-”, and is followed by a number signifying the order of discovery. If miRNAs only differ by one of two nucleotides, they are followed by an additional letter or number in their name. (Colpaert & Calore)

miRNAs can regulate gene expression by binding to different RNAs and changing the expression of proteins associated with them. A single piece of miRNA can regulate multiple mRNA targets, even if they are not perfectly complementary, proving its ability to target a wide range of cellular processes. (Seyhan)

miRNA can bind to sequences at the 3’ or 5’ untranslated regions of mRNA to dictate their translational ability (Ipsaro & Joshua-Tor). miRNAs can silence translation by binding to mRNA, blocking ribosomes from synthesizing proteins. Similarly, some miRNA can bind and induce the degradation of mRNA, halting the translation process. miRNAs can also overregulate gene expression by binding to mRNAs that encode repressor proteins. Without these repressors, genes can be freely expressed (O'Brien et al.,).

miRNAs have the ability to regulate gene expression post-transcriptionally by inhibiting protein synthesis. These nucleotides primarily do this by inhibiting mRNA, causing degradation of unnecessary proteins associated with them. miRNAs could even allow expression of beneficial genes by inhibiting repressor proteins with the same logic of gene silencing. The functions of miRNA are crucial to maintaining accurate expression of human genomes. However, in certain cases, dysregulation of miRNAs can cause normal gene silencing to go wrong. In disease scenarios, miRNAs could incorrectly degrade mRNAs, either inhibiting genes that should be active or oversuppressing genes to a harmful degree. In this section, gene silencing errors in a variety of human diseases will be explored and explained.

In cancer, cells have altered expressions of genes involved in cell proliferation, survival, and death. These differences in expression are due to changes in miRNAs. (Otmani et al.,)

In cancer, miRNAs are typically classified into two categories: oncomiR or TS (tumor suppressor). oncomiRs are overexpressed in individuals with cancer. They inhibit mRNAs of

tumor suppressor genes, allowing overgrowth of tumors. Oncogenes are mutated proto-oncogenes that normally control cell division. Oncogenes promote uncontrolled cell proliferation, which can contribute to tumor formation. TS miRNAs, which inhibit proto-oncogene mRNAs, are underexpressed in individuals with cancer. Thus, they lack regulated cell division mechanisms that oncogene mRNAs usually provide. (Otmani et al.,)

Some miRNAs can exhibit similar qualities even in the context of a different cancer cell. miR-21 is one of these examples. In glioblastoma, a type of cancer originating in the brain or spinal cord, miR-21 acts as a tumor promoter by affecting HNRPK and TAp63, which are key parts of the p53 tumor suppressor pathway. Similarly, in hepatocellular cancer in the liver, miR-21 acts as an oncomiR by excessively promoting cell proliferation (Meng et al.,).

miR-30 is another example of a miRNA involved in cancer. It acts as an oncomiR in pancreatic cancer by affecting CD133, a glycoprotein involved in cell signaling and promoting division of the cancerous cell (Tsukasa et al.,).

miR-21 is also a potential biomarker in cardiovascular diseases. In conditions like proliferative vascular disease, cardiac hypertrophy and heart failure, and ischemic heart disease, miR-21 has been dysregulated among test subjects. In recent studies, it was found that miR-21 was upregulated in cardiac fibroblasts in failing hearts. The upregulated miR-21 was shown to increase ERK-MAP kinase signaling pathways, leading to fibrosis, or scarring of tissue (Thum et al.,). miR-21 is also currently identified to target genes like PDCD4 which programs cell death, proving that miR-21 plays important roles in vascular smooth muscle cell proliferation and apoptosis, which are target processes associated with cardiovascular diseases. Thus, it is reasonable to hypothesize that the deregulation of miR-21 plays a crucial role in the development of cardiovascular diseases and their symptoms. (Cheng & Zhang).

While these findings are consistent in mouse models, the results have been ambiguous in humans. The unclear results could be because of differences in patients, diseases, and the detection methods of miR-21 (Cheng & Zhang).

In the nervous system, miR-132 plays a crucial role in differentiating neurons, managing their maturation, and regulating axon growth. miR-132 targets downstream molecules in biological pathways relating to these activities. The dysregulation of miR-132 is associated with neurodegenerative illnesses like Alzheimer's disease, Parkinson's disease, and epilepsy, as well as a wide range of psychiatric conditions like memory loss, depression, and schizophrenia (Qian et al.,).

Alzheimer's disease (AD) is the most prevalent form of dementia that is predicted to affect over 100 million people over the span of 30 years. In AD, synapses and neurons in the brain suffer from high rates of degradation in the basal forebrain, cortex, and hippocampus, causing significant cognitive decline. miRNA-132 indirectly influences the expression of genes such as APP and BACE1 through regulation of upstream signaling pathways (Qian et al.,).

Recently, in a study by Hon-Kit Andus Wong and his colleagues (Wong et al.), dysregulation of miR-132 was detected in the early onset of AD. Without the function of miR-132, the genes PTEN and FOXO3, which help regulate the cell cycle, will be expressed, and help trigger apoptosis (Qian et al.).

There are two main applications of miRNA therapeutics. miRNA may act as a replacement mimic of a tumor suppressor. This strategy induces loss of function in malignant targets, offering a novel approach to gene therapy. The miRNA replacement also represents an ability to explore tumor suppressors and their potential. However, the application of this method has yet to be clinically validated. Another more promising approach aims toward inhibiting oncogenic miRNAs through the use of antagonists called anti-miRs. These molecules are oligonucleotides that bind complementary to the oncogenic miRNA, leading to sequestration and degradation (Bader et al.).

Despite its potential, miRNA therapeutics face several challenges when trying to deliver these oligonucleotides *in vivo*. This is primarily due to enzyme degradation of these molecules during delivery. While solutions have been proposed to make chemical modifications to the miRNA to try to reduce risks of degradation, modifications may lead to off-target effects, affect the function of miRNA, or produce other toxicities (Rupaimoole,).

Introduction to Nanotechnology in Medicine

Nanotechnology refers to the study and application of technology at the nanosize scale with structures that have dimensions from 1 to 100 nanometers (10^{-9}m) (Wang).

Due to their small size, nanotechnology exhibits different biological properties, allowing new avenues of study and research (McNeil). Their customizable surfaces allow enhanced solubility and versatility, addressing challenges that arise during *in-vivo* delivery systems (Wang).

Altering the size, shape, or chemistry of nanoparticles allows them to be tailored to deliver specific types of molecules (Kulkarni & Feng). Nanoparticles need to rely on efficient cellular uptake and permeability in order to deliver accurate molecules to target cells and bypass obstacles (Catalanotto). Additionally, different types of destinations that miRNAs are delivered to require different sizes of nanoparticles, and thus, the customizability of nanoparticles allows for optimized benefits (Hoshyar).

The use of nanocarriers in the delivery, release, and treatment of miRNA provides the positive effects of stability, bioavailability, and tissue specificity. Some important nanocarriers commonly used in miRNA applications include inorganic nanoparticles, polymeric nanoparticles, lipid nanoparticles, and carbon-based nanomaterials (Wang).

There are three categories of methods to synthesize nanoparticles: biological, chemical, and physical. The physical approach is often referred to as the “top-down” approach, which refers to the breakdown of larger materials into nanosized particles. The chemical and biological methods could also be referred to as the bottom-up approach, which explains the method that accumulates small molecules to create a nanoparticle (Altammar).

Among top-down approaches of nanoparticle synthesis, ball milling, lithography, and laser ablation are the most common. (Altammar)

Ball milling is a process that uses hard metal balls in various mills, usually planetary or shaker mills, to break down materials in a high-impact process to a nanoparticle size. Ball-milled carbon nanomaterials, a unique class of nanoparticles that are made using the ball milling technique, often help with energy storage, energy conversion, and ecological restoration purposes (Yadav, Altammar).

Lithography is the process that utilizes a concentrated beam of light or electrons to carve out nanoparticles. Within lithography, there are two primary categories: mask and maskless. Mask lithography uses a physical barrier to selectively allow light to pass through and create precise nanoscale patterns. Maskless lithography uses software-controlled systems to dictate shape instead of using a physical barrier, creating a more flexible option. Unwanted particles that are left after the carving process are then washed away, leaving behind nanoscale patterns (Altammar).

Laser ablation refers to the process in which a laser beam vaporizes a material. As the vapor cools, small nanoparticles are created as a result. This method has the benefits of being environmentally friendly and can produce a wide range of nanoparticles (Baig et al.,; Altammar).

Bottom-up approaches refer to the combination of small atoms to create nanosized structured particles (Altammar). Within this category, sol-gel, chemical vapor deposition (CVD), co-precipitation, and self-assembly are common.

In the sol-gel method, desired metals inside a solution are condensed, hydrolyzed, and thermally decomposed to create a sol solution. Then, the metals inside the solution combine together, increasing the viscosity and creating a gel. It takes time for the solvent to be removed and for the nanoparticles to combine into a solid mass. Lastly, to separate the nanoparticles, the chemical ingredients in the solution are heated. The ease in creating complex structures and the quality of material are some of the advantages of the sol-gel technique (Altammar).

In the chemical vapor deposition process (CVD), chemical gases are sent into a chamber. Within that chamber, there is a heated substrate surface made of glass or silicon. As the vaporized chemicals engage with the substrate surface, a thin film will be created. This thin film builds up to nanoparticle solids. The CVD process has the benefits of creating pure, homogenous nanoparticles and has shown great success in creating two-dimensional structures (Altammar).

In the chemical co-precipitation technique, chemical solutions are combined in water. Then a base solution, usually made of ammonium, is added to the solution, changing its pH and other chemical properties. This change allows for molecules to clump up and create desired nanoparticles. This method is often used due to its relative simplicity and economic efficiency. It can also easily create large quantities of nanoparticles. (Altammar).

Self-assembly is the process by which individual atoms or molecules arrange together to create ordered structures from local interactions amongst themselves. This process could work with both inorganic and organic structures via non-covalent bonds. Some examples of self-assembly are the construction of cell membranes with phospholipid bilayers, the helical

structure of DNA, or the folding of polypeptide chains in protein formation. Self-assembled nanomaterials have capabilities in a wide variety of applications. Nanotechnology in medicine requires specific nanostructures with built-in desired properties (Cui et al.,). Self-assembly allows nanotechnology to possess the same parameters as the natural material by resembling their natural 3-D structure. Self-assembly has the benefits of being economically efficient and easy to use. Additionally, the technique has a wide versatility as different structures could be achieved by varying the monomer, utilizing its chemical properties to fit the desired use (Yadav).

Different nanoparticles have been employed for delivering therapeutic molecules to targeted sites without being degraded. The nanoparticles possess distinct functional properties based on their composition, surface characteristics, and size.

Lipid nanoparticles (LNPs) are vesicles composed of lipids. This category of nanoparticles has shown the most success in the delivery of nucleic acids. Recently, mRNA-based vaccines, such as those developed by Pfizer and Moderna for COVID-19, have helped gain massive attention and hope for their potential. LNPs are made up of four different components: ionizable lipid, PEGylated lipid, phospholipid, and cholesterol. Its phospholipid outer shell creates a hydrophilic exterior and a hydrophobic interior. This dual property creates an optimal delivery system for nucleic acids, as its negatively charged phosphate group demonstrates hydrophilic qualities, allowing miRNA to be encapsulated within the vesicle and protecting it from nucleases in the body that degrade during delivery (Jung et al.,).

Polymeric nanoparticles, ranging from 1 to 1000 nm, are made from biodegradable plastics or natural materials like PLGA, a type of synthetic polymer, and chitosan, a natural polymer from shellfish. Polymeric NPs can be classified as nanocapsules or nanospheres based on their morphological structure. Nanocapsules have an oily core that encapsulates the drug and a polymeric shell that determines the release of the drug. Nanospheres have a network that allows drugs to be absorbed inside or onto the surface. In the creation of nanoparticles, polyplex complexes are formed by combining a polymer and a nucleic acid. The difference in charges of a nucleic acid, which is usually negative, and a polymer, which is often positive, creates an attraction that allows the miRNA to be encapsulated in a compact form, or coated on the outside (Zielińska et al.,). Advantages of polymeric nanoparticles include their ability to control the release of the drug and their ability to protect them from biological degradation. Polymeric NPs are beneficial as they are safe for human use and have the ability to improve the bioavailability of loaded drugs. The use of natural polymers like chitosan, PLGA, and others has had success in drug deliveries due to the materials' biodegradable and biocompatible qualities (Zielińska et al.,).

Due to their inorganic nature, metallic nanoparticles have unique characteristics like sensitivity to light and electrons. Additionally, their magnetic properties allow for easy manipulation of shape and size (Altammar). For example, they can carry miRNA molecules through chemical covalent bonding. However, since they are inorganic, they have risks of inflammation and toxicity with prolonged exposure and are not biodegradable (Bhatti et al.,).

Dendrimers are symmetrical nanoparticles where a small group of atoms is surrounded by branches called dendrons. Dendrimers have positively charged surface functional groups that

could bind to the negative phosphate on miRNA. Additionally, they can be functionalized with ligands like antibodies and peptides that guide them to the target cell, reducing off-target effects. Their three-dimensional branched structure allows for a variety of drugs to be carried, forming drug conjugates. By applying different dendritic patterns, synergistic effects could be achieved. The combination of compounds on individual dendrons has more benefits than single molecules. Dendrimers have most commonly been used to enhance absorption and cellular uptake of drugs in therapeutics (Mittal et al.,).

Therapeutic Strategies with miRNA

miRNAs can be used as a potential therapeutic strategy to control regulation of diseases. miRNA expression can be controlled by either delivering beneficial miRNAs through mimics or by inhibiting the harmful miRNA with anti-miRs. These approaches, which form the basis of current therapeutic developments in this field, are discussed in this section.

In miRNA therapeutics, two primary strategies are commonly employed. The first strategy involves restoring the function of beneficial miRNAs using miRNA mimics. In diseases with underexpressed miRNA, malignant proteins are overexpressed as they are no longer inhibited. miRNAs are restored in this strategy in order to re-establish a loss-of-function of harmful proteins. In simpler words, this strategy aims to replace tumor suppressors and mimic disease blocking genes (Bader et al.,).

Although disease blocking proteins like tumor suppressors can also be restored via DNA delivery, miRNA mimics encounter much fewer delivery barriers. Since miRNA mimics are smaller, they are able to undergo their function once delivered into the cytoplasm of their target cell, unlike DNA, which has to go into the nucleus. Additionally, another advantage of miRNA mimics is that they have the same sequence identity as their natural depleted counterpart, making unwanted off-target effects unlikely (Bader et al.,).

miR-34s have a large variety of functions due to their broad range of mRNA targets across different cells. miR-34 is responsible for targeting cell survival pathways like the MAPK pathway. Thus, dysregulation of miR-34 has often been seen in cancer patients. miR-34 expression is downregulated compared to normal tissues in breast cancer and ovarian cancer. In order to restore normal expression of these genes, delivery of miR-34 mimics can be transported to target cells and restore normal function of cell survival pathways as a therapeutic method (Fu et al.,).

In addition to miR-34, dysregulation of miR-126 poses issues in cancer and cardiovascular disease. Currently, scientists hypothesize that the downregulation of miR-126 contributes to a variety of vascular diseases and cancers, while upregulation is beneficial in instances such as coronary artery disease. miR-126 has multiple different target genes and is able to dictate its function based on cellular conditions. miR-126 is highly expressed in endothelial cells, where it helps in angiogenesis (i.e., formation of new blood vessels) (Yu et al.,).

To restore high concentrations of miR-126 in diseased patients, miR-126 mimics could be delivered via nanoparticles to support its therapeutic functions.

Another therapeutic strategy involves inhibiting harmful miRNAs using antagomiRs (anti-miRNAs). These oligonucleotides have sequences complementary to harmful endogenous miRNAs. When anti-miRNA binds to miRNA, it prevents their processing by the RISC (RNA-induced silencing complex), thereby inhibiting their gene-silencing function. antimiRs could also degrade miRNA by binding to them. The goal of this technique is to inhibit overexpressed miRNAs that may be silencing required beneficial genes (Bader et al.,).

miR-21 is an example of an upregulated miRNA found in cancers like glioma, breast cancer, and colorectal cancer. It targets and inhibits the expression of genes like PDCD4 (programmed cell death protein 4), which play huge roles in cell cycles. By inhibiting programmed cell death, uncontrolled apoptosis could occur, leading to unregulated cell survival, causing the formation of tumors or cancer progression. Researchers are investigating the delivery of anti-miR-21 oligonucleotides to inhibit miR-21 and restore genes like PDCD4 (Bautista-Sánchez et al.,). However, despite its capabilities in therapeutics, anti-miR-21 has been unable to adapt to the tumor microenvironment of many cancerous cells. Thus, inhibiting miR-21 has not been successful across all cancer types (Bautista-Sánchez et al.,).

Inflammation can often be associated with the development and progression of cancers by triggering cancerous mutations (Singh et al.,). miR-155 acts as a promoter for inflammation. The overexpression of miR-155 has been involved in cell growth, migration, and many more pathways involved in diseases such as lung cancer, kidney cancer, and breast cancer. In a study by Tili and her colleagues, miR-155 has been shown to cause tumor progression by targeting tumor suppressor genes (Tili et al.,). With miR-155's harmful role, reducing its expression with anti-miR-155 emerges as a beneficial treatment of inflammatory cancers (Hu et al.,).

Integrating Nanotechnology and microRNA for Precision Therapy

miRNA therapeutic efficiency is compromised by the rapid degradation, poor targeting, and inefficient drug uptake of miRNA. To overcome these issues, nanoparticles loaded with miRNA are introduced to protect these nucleotides from degradation inside the body and minimize their risks (Lee et al.,).

The stability of miRNA structures is compromised during delivery. For example, unprotected miRNA with unmodified OH groups without nanocarriers has the possibility of being degraded rapidly by nucleotide-cleaving enzymes (nucleases) such as serum RNase A-type nucleases in the blood. Unprotected miRNAs could also be cleared by renal excretion in the body's normal maintenance and homeostasis (Segal & Slack).

miRNAs are able to target many different biological pathways with various unique functions. Their capabilities of imperfectly binding to various target mRNAs open up a wide range of different proteins they could regulate. But this binding behavior allows for undesirable gene silencing of non-targets. Without protection, miRNAs have a high possibility of initiating off-target side effects (Segal & Slack).

Lastly, cellular uptake of therapeutic miRNA poses a severe problem, especially in cancer treatments. This is because tumor-induced angiogenesis, or excess of blood vessel formation,

allows elimination of the drug through nuclease degradation or immune system mechanisms (Gupta & Qin). As a result, attaining sufficient miRNA for delivery through endocytosis becomes a challenge (Grixti et al.).

Due to the challenges of miRNA delivery, researchers have explored miRNA-loaded nanoparticles. As a result of their unique biological and chemical properties, nanoparticles are able to shield miRNA from degradation and assist in targeted delivery with controlled release. Nanoparticles have unique ways of loading miRNA as well as creating targeted delivery routes. The use of nanoparticles allows for the benefits of miRNA therapeutics while simultaneously reducing their limitations (Lee et al.).

Nanoparticles have structures that allow miRNAs to be safely loaded and protected during the delivery process.

Encapsulation refers to the method of loading miRNA molecules into the internal voids of a nanoparticle. This process could be assisted by hydrogen bonding and electrostatic forces of miRNA nucleic acids. Hydrogen bonds hold the structure of miRNA and facilitate its interaction with the nanoparticle carrier, making the encapsulation stable. The nanoparticle delivers the miRNA to its desired site for release (Deka Dey et al.). Dendrimers are attractive candidates for encapsulation methods of loading miRNAs. This is because of their highly branched polymeric structure, creating empty internal voids. These internal voids allow space for miRNA to be encapsulated and protected from nucleases in the body during delivery (Zare et al.).

Electrostatic binding is utilized when cationic carriers, like lipids and polymers, form electrostatic interactions with nucleic acids due to opposite charges during drug loading. The negative phosphate backbone of the miRNA attracts positively charged groups on polymers, allowing miRNAs to bind to the surfaces of nanocarriers with stability. In addition to organic carriers, some inorganic nanoparticles like gold can also employ electrostatic surface binding loading techniques. Gold can be functionalized with amino groups that have a positive component. The positive amino groups can create an electrostatic bond with negative phosphate groups of miRNA during loading (Moraes et al.).

miRNA loaded nanoparticles delivery have varying degrees of success depending on the unique biological capabilities of the patient and disease it is targeting.

Intravenous Delivery Intravenous delivery is the method by which drugs are administered directly into the vein. This method is especially helpful for miRNA molecules since they have low molecular weights and are soluble in water. After injecting the drug with intravenous delivery methods, miRNA mimics and anti-miRNA accumulate in the liver and kidney, while brain exposure is limited by the blood-brain barrier restricting entry. In the delivery process, modified miRNAs are capable of remaining stable in tissues long after their administration, creating prolonged therapeutic benefits while unmodified counterparts suffer from rapid degradation (Chen et al.).

Local delivery administers the drug directly to the target area. Its ability to quickly become active and administer therapeutic benefits is efficient. For example, in topical delivery

for skin disease, administering the drug directly on the target region allows for reduced side effects and focused delivery. Local delivery, however, has a limited role in therapy since it only targets readily accessible areas (Chen et al.,).

One approach to reduce off-target effects of miRNA delivery is to utilize surface modifications of nanoparticles that interact with target cells. An example of a possible modification is PEGylation. PEGylation attaches a polyethylene glycol, changing the chemical properties of a molecule, and reducing interactions with off-target cells and immune effects. As a result, it makes it less likely for nucleases to break down the drugs being delivered (Veronese & Mero). In addition, target ligands can be incorporated into surfaces of nanoparticles. These target ligands bind to specific tissues, reducing the possibility of off-target interactions with other cells. Size and shape of nanoparticles could also be modified to allow more efficient delivery. For example, smaller nanoparticles are more likely to pass through renal pathways and thus reduce chances of accumulation (Telkoparan-Akillilar et al.,).

Unlike other small molecules, RNA is too large and too hydrophilic to be able to passively diffuse through the lipid bilayer of the cell membrane. Thus, it relies on the process of endocytosis assisted by molecules called endosomes. However, endosomes capture 99% of RNA deliveries, only allowing 1% to enter the cytoplasm for therapeutics (Dowdy).

To solve this issue, LNPs are used to adequately deliver miRNA past the barriers of endosomes into the cytoplasm for release. LNPs are made from four components: an ionizable lipid (IL), a phospholipid helper lipid, cholesterol, and a polyethylene glycol lipid. The ionizable lipid is the primary component in helping miRNA escape the endosome. The IL becomes cationic and water-soluble when it interacts with the acidic pH inside endosomes. As a result, the cationic lipids of the LNP and the anionic lipids of the endosome interact, eventually triggering the endosome's membrane to break apart and releasing miRNA into the cytoplasm (Melina, Wagner)

pH-sensitive particles swell or dissolve as a result of interactions with pH, allowing them to release miRNA at the desired destination. LNPs are not the only type of pH-sensitive nanoparticle that assists with the accurate delivery of miRNA. Particles made of metal or polymers are also capable of exhibiting this characteristic. For example, nanocarriers made of many carboxylic acid groups stay compact at acidic pH. However, once the pH becomes more neutral, these carboxylic groups lose a proton, loosening the structure and breaking the polymer apart to release the miRNA. Nanocarriers are able to respond to the specific pH changes at their target destination, causing release of miRNA at the accurate spot without off-target effects (Karimi et al.,).

As mentioned previously, miR-34 targets a variety of pathways regarding cell survival and proliferation, like the MAPK pathway. For that reason, miR-34 has become a reasonable biomarker for many cancer diseases. Delivering miR-34 could act as a treatment for cancer diseases by restoring accurate regulation of these pathways (Fu et al.,).

Liposomes, nanoparticles with lipid bilayers, have been shown to be capable of delivering miR-34 mimics. They can shield the miRNA from nuclease degradation through electrostatic binding, assist with their delivery into cells via endocytosis, and allow controlled release. In addition, polymeric nanoparticles and other types of lipid nanoparticles are able to provide benefits of surface modifications, assisting in target delivery and cellular uptake to improve therapeutic outcomes. The approach in delivering miR-34 to treat cancer has shown promise in studies, thanks to the properties and potential of nanoparticle delivery systems (Fu et al.,).

In addition, exosome nanocarriers loaded with miR-132 have been explored preclinically in the treatment of ischemic heart diseases, where blood arteries are narrowed, reducing the flow of oxygen (Board on the Health of Select Populations & Committee on Social Security Cardiovascular Disability Criteria). miR-132 can promote angiogenesis by downregulating the gene *RASA1*. Exosomes are able to protect miR-132 from degradation due to their membrane-bound structure and assist in delivery, making them attractive options in miRNA therapeutics (Ma et al.,).

Challenges and Future Directions

While nanotechnology provides many benefits, there are significant consequences to using it as a delivery mechanism in miRNA therapeutics. Nanotechnology faces challenges in both biological and clinical aspects such as toxic side effects to the human body and difficulty in approval as a drug due to their complex properties. However, besides these challenges, the use of nanotechnology with miRNA allows for promising future opportunities such as theranostics and improvements with more intricate personalized medicine treatments.

Nanoparticles have been reported in different studies for potential toxicity risks in the human body. For example, inorganic metal-based nanoparticles have side effects like neurotoxicity due to long circulation time and off-target accumulation. In addition, since this treatment method is still in its early stages of development, hazards to humans are not yet fully characterized, necessitating caution (Zhang et al.,).

Nanotechnology and miRNA therapeutics have also run into problems clinically. This mode of treatment has struggled in development and certification. Nanoparticle specificity makes it difficult to evaluate their potential and safety. Nanoparticles are complex three-dimensional structures. Each characteristic of its design has specific chemical characteristics and behaviors. As a result, minor variations in its structure can cause negative consequences in behavior, safety, and efficiency, creating the struggle of having to study each nanoparticle thoroughly (Desai).

Since nanoparticles are highly specific to their intended use, there are no existing models that could sustain all the diverse functions of nanotechnology. Thus, the development of these tools relies on heavy testing and evidence, taking up a lot of time (Desai).

The futures of precision medicine allow personalized treatments to target an individual's unique biological environment. This scope of medicine allows for the creation of more effective treatments and early disease prevention.

The potential of precision medicine continues to develop with improvements in nanocarriers that respond to environmental conditions. Nanocarriers have been engineered with abilities to respond to both external stimuli, including light and magnetic field changes, and internal stimuli, including pH and enzyme functions. This potential has become powerful in cancer treatments by responding to the tumor environment and providing controlled and specific drug delivery (Mi).

Specifically, pH-sensitive nanocarriers have been utilized due to the acidic nature of organelles. Since the pH of a healthy cytoplasm is around 7.0-7.4 and tumor environments range from 6.5- 6.8, sensitivities to pH changes could provide accurate drug release in desired target destinations. Currently, scientists have begun studying pH-sensitive nanocarriers like liposomes, dendrimers, CaCO₃ nanoparticles, and many more (Mi).

Additionally, scientists are looking to a new branch of personalized medicine. Theranostics combine diagnostics and therapeutics to detect and treat diseases at the same time. This field also allows therapeutics to be specific to what was found in the disease's progression. Since the approval by the U.S Food and Drug Administration (FDA) of lutetium 177, an isotope used to treat prostate cancer, the field of theranostics has gained support as a potential improvement to precision medicine (Burkett et al.).

Theranostics have been improved by developments in AI, enabling automated medical image analysis and detecting imperceptible features like tumor radiosensitivity, offering a more detailed evaluation of disease progression. The developments of AI in theranostics are continuing to provide patients with more appropriate treatment plans (Belge et al.).

Conclusion

miRNAs have promising therapeutic potential due to their ability to regulate gene expression at the post-translational level. They hold potential to enhance beneficial gene expression by using miRNA mimics and suppress harmful genes with anti-miRNAs. While miRNAs provide \valuable therapeutic strategies, they often face challenges during their delivery. They are prone to nuclease degradation during transport, exhibit reduced cellular uptake in treatment, and can cause off-target side effects with their imperfect binding nature to mRNA. Consequently, new novel nanocarrier systems have emerged as potential solutions to these issues. Nanocarriers possess specific biological characteristics that protect miRNA during *in vivo* delivery and can also be engineered with surface modifications to assist in targeted delivery. The integration of nanotechnology into miRNA therapeutics addresses several challenges with delivery, providing potential for improved treatment outcomes. Together, they can help shape the future of precision therapy and broaden the application across a wide range of diseases.

Works Cited

- Altammar, K. A. "A Review on Nanoparticles: Characteristics, Synthesis, Applications, and Challenges." *Frontiers in Microbiology*, vol. 14, 2023, p. 1155622. <https://doi.org/10.3389/fmicb.2023.1155622>.
- Bader, A. G., D. Brown, and M. Winkler. "The Promise of MicroRNA Replacement Therapy." *Cancer Research*, vol. 70, no. 18, 2010, pp. 7027–7030. <https://doi.org/10.1158/0008-5472.CAN-10-2010>.
- Baig, N., I. Kammakakam, and W. Falath. "Nanomaterials: A Review of Synthesis Methods, Properties, Recent Progress, and Challenges." *Materials Advances*, vol. 2, no. 6, 2021, pp. 1821–1871.
- Bautista-Sánchez, D., et al. "The Promising Role of miR-21 as a Cancer Biomarker and Its Importance in RNA-Based Therapeutics." *Molecular Therapy – Nucleic Acids*, vol. 20, 2020, pp. 409–420. <https://doi.org/10.1016/j.omtn.2020.03.003>.
- Belge Bilgin, G., et al. "Theranostics and Artificial Intelligence: New Frontiers in Personalized Medicine." *Theranostics*, vol. 14, no. 6, 2024, pp. 2367–2378. <https://doi.org/10.7150/thno.94788>.
- Bhatti, R., et al. "Inorganic Nanoparticles: Toxic Effects, Mechanisms of Cytotoxicity and Phytochemical Interactions." *Advanced Pharmaceutical Bulletin*, vol. 12, no. 4, 2022, pp. 757–762. <https://doi.org/10.34172/apb.2022.077>.
- Board on the Health of Select Populations, and Committee on Social Security Cardiovascular Disability Criteria. *Cardiovascular Disability: Updating the Social Security Listings*. National Academies Press, 2010.
- Burkett, B. J., et al. "A Review of Theranostics: Perspectives on Emerging Approaches and Clinical Advancements." *Radiology: Imaging Cancer*, vol. 5, no. 4, 2023, e220157.
- Catalanotto, C., C. Cogoni, and G. Zardo. "MicroRNA in Control of Gene Expression: An Overview of Nuclear Functions." *International Journal of Molecular Sciences*, vol. 17, no. 10, 2016, p. 1712. <https://doi.org/10.3390/ijms17101712>.
- Chen, Y., D. Y. Gao, and L. Huang. "In Vivo Delivery of miRNAs for Cancer Therapy: Challenges and Strategies." *Advanced Drug Delivery Reviews*, vol. 81, 2015, pp. 128–141. <https://doi.org/10.1016/j.addr.2014.05.009>.
- Cheng, Y., and C. Zhang. "MicroRNA-21 in Cardiovascular Disease." *Journal of Cardiovascular Translational Research*, vol. 3, no. 3, 2010, pp. 251–255. <https://doi.org/10.1007/s12265-010-9169-7>.
- Colpaert, R. M. W., and M. Calore. "MicroRNAs in Cardiac Diseases." *Cells*, vol. 8, no. 7, 2019, p. 737. <https://doi.org/10.3390/cells8070737>.
- Desai, N. "Challenges in Development of Nanoparticle-Based Therapeutics." *The AAPS Journal*, vol. 14, no. 2, 2012, pp. 282–295. <https://doi.org/10.1208/s12248-012-9339-4>.
- Dowdy, S. F. "Endosomal Escape of RNA Therapeutics: How Do We Solve This Rate-Limiting Problem?" *RNA*, vol. 29, no. 4, 2023, pp. 396–401. <https://doi.org/10.1261/rna.079507.122>.

- Fu, J., S. Imani, M. Y. Wu, and R. C. Wu. "MicroRNA-34 Family in Cancers: Role, Mechanism, and Therapeutic Potential." *Cancers*, vol. 15, no. 19, 2023, p. 4723. <https://doi.org/10.3390/cancers15194723>.
- Grixti, J. M., D. Ayers, and P. J. R. Day. "An Analysis of Mechanisms for Cellular Uptake of miRNAs to Enhance Drug Delivery and Efficacy in Cancer Chemoresistance." *Non-Coding RNA*, vol. 7, no. 2, 2021, p. 27. <https://doi.org/10.3390/ncrna7020027>.
- Gupta, M. K., and R. Y. Qin. "Mechanism and Its Regulation of Tumor-Induced Angiogenesis." *World Journal of Gastroenterology*, vol. 9, no. 6, 2003, pp. 1144–1155. <https://doi.org/10.3748/wjg.v9.i6.1144>.
- Hoshyar, N., S. Gray, H. Han, and G. Bao. "The Effect of Nanoparticle Size on In Vivo Pharmacokinetics and Cellular Interaction." *Nanomedicine (London, England)*, vol. 11, no. 6, 2016, pp. 673–692. <https://doi.org/10.2217/nmm.16.5>.
- Hu, J., et al. "miR-155: An Important Role in Inflammation Response." *Journal of Immunology Research*, 2022, 7437281. <https://doi.org/10.1155/2022/7437281>.
- Ipsaro, J. J., and L. Joshua-Tor. "From Guide to Target: Molecular Insights into Eukaryotic RNA-Interference Machinery." *Nature Structural & Molecular Biology*, vol. 22, no. 1, 2015, pp. 20–28. <https://doi.org/10.1038/nsmb.2931>.
- Jung, H. N., et al. "Lipid Nanoparticles for Delivery of RNA Therapeutics: Current Status and the Role of In Vivo Imaging." *Theranostics*, vol. 12, no. 17, 2022, pp. 7509–7531. <https://doi.org/10.7150/thno.77259>.
- Karimi, M., et al. "pH-Sensitive Stimulus-Responsive Nanocarriers for Targeted Delivery of Therapeutic Agents." *Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology*, vol. 8, no. 5, 2016, pp. 696–716. <https://doi.org/10.1002/wnan.1389>.
- Kulkarni, S. A., and S. S. Feng. "Effects of Particle Size and Surface Modification on Cellular Uptake and Biodistribution of Polymeric Nanoparticles for Drug Delivery." *Pharmaceutical Research*, vol. 30, no. 10, 2013, pp. 2512–2522. <https://doi.org/10.1007/s11095-012-0958-3>.
- Lee, S. W. L., et al. "MicroRNA Delivery through Nanoparticles." *Journal of Controlled Release*, vol. 313, 2019, pp. 80–95. <https://doi.org/10.1016/j.jconrel.2019.10.007>.
- Lee, T. I., and R. A. Young. "Transcriptional Regulation and Its Misregulation in Disease." *Cell*, vol. 152, no. 6, 2013, pp. 1237–1251. <https://doi.org/10.1016/j.cell.2013.02.014>.
- Ma, T., et al. "MicroRNA-132, Delivered by Mesenchymal Stem Cell-Derived Exosomes, Promote Angiogenesis in Myocardial Infarction." *Stem Cells International*, 2018, 3290372. <https://doi.org/10.1155/2018/3290372>.
- McNeil, S. E. "Nanotechnology for the Biologist." *Journal of Leukocyte Biology*, vol. 78, no. 3, 2005, pp. 585–594. <https://doi.org/10.1189/jlb.0205074>.
- Meng, F., et al. "MicroRNA-21 Regulates Expression of the PTEN Tumor Suppressor Gene in Human Hepatocellular Cancer." *Gastroenterology*, vol. 133, no. 2, 2007, pp. 647–658. <https://doi.org/10.1053/j.gastro.2007.05.022>.

- Melina, G., and E. Wagner. "Strategies and Mechanisms for Endosomal Escape of Therapeutic Nucleic Acids." *Current Opinion in Chemical Biology*, vol. 81, 2024, 102506. <https://doi.org/10.1016/j.cbpa.2024.102506>.
- Mi, P. "Stimuli-Responsive Nanocarriers for Drug Delivery, Tumor Imaging, Therapy and Theranostics." *Theranostics*, vol. 10, no. 10, 2020, pp. 4557–4588. <https://doi.org/10.7150/thno.38069>.
- Mittal, P., et al. "Dendrimers: A New Race of Pharmaceutical Nanocarriers." *BioMed Research International*, 2021, 8844030. <https://doi.org/10.1155/2021/8844030>.
- Moraes, F. C., C. Pichon, D. Letourneur, and F. Chaubet. "miRNA Delivery by Nanosystems: State of the Art and Perspectives." *Pharmaceutics*, vol. 13, no. 11, 2021, p. 1901. <https://doi.org/10.3390/pharmaceutics13111901>.
- O'Brien, J., H. Hayder, Y. Zayed, and C. Peng. "Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation." *Frontiers in Endocrinology*, vol. 9, 2018, p. 402. <https://doi.org/10.3389/fendo.2018.00402>.
- Otmani, K., R. Rouas, and P. Lewalle. "OncomiRs as Noncoding RNAs Having Functions in Cancer: Their Role in Immune Suppression and Clinical Implications." *Frontiers in Immunology*, vol. 13, 2022, 913951. <https://doi.org/10.3389/fimmu.2022.913951>.
- Qian, Y., et al. "Advances in Roles of miR-132 in the Nervous System." *Frontiers in Pharmacology*, vol. 8, 2017, p. 770. <https://doi.org/10.3389/fphar.2017.00770>.
- Rupaimoole, R., H. D. Han, G. Lopez-Berestein, and A. K. Sood. "MicroRNA Therapeutics: Principles, Expectations, and Challenges." *Chinese Journal of Cancer*, vol. 30, no. 6, 2011, pp. 368–370. <https://doi.org/10.5732/cjc.011.10186>.
- Segal, M., and F. J. Slack. "Challenges Identifying Efficacious miRNA Therapeutics for Cancer." *Expert Opinion on Drug Discovery*, vol. 15, no. 9, 2020, pp. 987–992. <https://doi.org/10.1080/17460441.2020.1765770>.
- Seyhan, A. A. "Trials and Tribulations of MicroRNA Therapeutics." *International Journal of Molecular Sciences*, vol. 25, no. 3, 2024, p. 1469. <https://doi.org/10.3390/ijms25031469>.
- Singh, N., et al. "Inflammation and Cancer." *Annals of African Medicine*, vol. 18, no. 3, 2019, pp. 121–126. https://doi.org/10.4103/aam.aam_56_18.
- Telkoparan-Akillilar, P., S. Chichiarelli, P. Tucci, and L. Saso. "Integration of MicroRNAs with Nanomedicine: Tumor Targeting and Therapeutic Approaches." *Frontiers in Cell and Developmental Biology*, vol. 13, 2025, 1569101. <https://doi.org/10.3389/fcell.2025.1569101>.
- Thum, T., et al. "MicroRNA-21 Contributes to Myocardial Disease by Stimulating MAP Kinase Signalling in Fibroblasts." *Nature*, vol. 456, no. 7224, 2008, pp. 980–984. <https://doi.org/10.1038/nature07511>.
- Tili, E., et al. "Mutator Activity Induced by MicroRNA-155 (miR-155) Links Inflammation and Cancer." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 12, 2011, pp. 4908–4913. <https://doi.org/10.1073/pnas.1101795108>.

- Tsukasa, K., et al. “miR-30 Family Promotes Migratory and Invasive Abilities in CD133(+) Pancreatic Cancer Stem-like Cells.” *Human Cell*, vol. 29, no. 3, 2016, pp. 130–137. <https://doi.org/10.1007/s13577-016-0137-7>.
- Veronese, F. M., and A. Mero. “The Impact of PEGylation on Biological Therapies.” *BioDrugs*, vol. 22, no. 5, 2008, pp. 315–329. <https://doi.org/10.2165/00063030-200822050-00004>.
- Wang, H. “A Review of Nanotechnology in microRNA Detection and Drug Delivery.” *Cells*, vol. 13, no. 15, 2024, p. 1277. <https://doi.org/10.3390/cells13151277>.
- Wong, H. K., et al. “De-Repression of FOXO3a Death Axis by MicroRNA-132 and -212 Causes Neuronal Apoptosis in Alzheimer's Disease.” *Human Molecular Genetics*, vol. 22, no. 15, 2013, pp. 3077–3092. <https://doi.org/10.1093/hmg/ddt164>.
- Yadav, S., A. K. Sharma, and P. Kumar. “Nanoscale Self-Assembly for Therapeutic Delivery.” *Frontiers in Bioengineering and Biotechnology*, vol. 8, 2020, p. 127. <https://doi.org/10.3389/fbioe.2020.00127>.
- Yadav, T. P., R. M. Yadav, and D. P. Singh. “Mechanical Milling: A Top Down Approach for the Synthesis of Nanomaterials and Nanocomposites.” *Nanoscience and Nanotechnology*, vol. 2, no. 3, 2012, pp. 22–48.
- Yu, B., Y. Jiang, X. Wang, and S. Wang. “An Integrated Hypothesis for miR-126 in Vascular Disease.” *Medical Research Archives*, vol. 8, no. 5, 2020, p. 2133. <https://doi.org/10.18103/mra.v8i5.2133>.
- Zare, M., et al. “Encapsulation of miRNA and siRNA into Nanomaterials for Cancer Therapeutics.” *Pharmaceutics*, vol. 14, no. 8, 2022, p. 1620. <https://doi.org/10.3390/pharmaceutics14081620>.
- Zhang, N., G. Xiong, and Z. Liu. “Toxicity of Metal-Based Nanoparticles: Challenges in the Nano Era.” *Frontiers in Bioengineering and Biotechnology*, vol. 10, 2022, 1001572. <https://doi.org/10.3389/fbioe.2022.1001572>.
- Zielińska, A., et al. “Polymeric Nanoparticles: Production, Characterization, Toxicology and Ecotoxicology.” *Molecules*, vol. 25, no. 16, 2020, p. 3731. <https://doi.org/10.3390/molecules25163731>

Acknowledgements

Huge thank you to my mentor, Muhammad Hassan Sarfraz, from Oxford University for his guidance and mentorship over the past few months. In addition, thank you to Lumiere Education for this wonderful opportunity to continue exploring my interests.

Reliable Gaze-Driven Medical Communication Boards Using Consistency Loss with Conjugate Gaze By JoonWoo Park and Hyunggyu Kim

ABSTRACT

Quadriplegia is a condition characterized by the partial or total loss of use of all four limbs and the torso. It is typically caused by an injury to the spinal cord at the cervical (neck) level or by neurological conditions affecting the spinal cord or brain. This condition severely impacts an individual's ability to speak or use hand gestures. To overcome these challenges, quadriplegic patients often rely on communication boards, such as picture or alphabet boards. These boards allow patients to express themselves by pointing to symbols, words, or letters using methods like direct selection. However, traditional communication boards can be cumbersome to use, as they require an assistant to guide the patient through the options one by one, making the process slow and inefficient. To address this problem, we propose a gaze estimation-based communication board that utilizes gaze vectors to select items. The system uses input from both eyes to predict the direction of the patient's gaze. To enhance accuracy, we introduce a conjugate gaze ability-based loss function. Our proposed system achieved an 8.8-degree error on a public gaze estimation dataset. We have implemented this system in real-world scenarios to demonstrate its efficiency and feasibility.

Introduction

Quadriplegia is a type of paralysis that affects all four limbs and the torso, usually resulting from a spinal cord injury. The spinal cord is a critical part of the central nervous system, functioning as a conduit for messages between the brain and the rest of the body. Quadriplegia typically results from a severe injury to the cervical (neck) region of the spinal cord, disrupting the transmission of nerve signals below the injury site and leading to paralysis of the limbs and trunk. Consequently, individuals with quadriplegia cannot voluntarily move their arms, legs, or torso.

However, eye movements are largely controlled by cranial nerves, which originate directly from the brain rather than the spinal cord. These nerves include the oculomotor (III), trochlear (IV), and abducens (VI) nerves, which control the muscles responsible for eye movements. Since these cranial nerves are above the level of the spinal cord injury, they remain intact and functional, allowing individuals with quadriplegia to move their eyes even though the rest of their body is paralyzed.

Traditionally, people with quadriplegia use a communication board to communicate with medical assistants or others. A communication board is a sheet of paper with symbols, pictures, and photos used as an alternative communication method. A person, often a nurse, points to each symbol one by one until the patient responds with their eyes. This method is very impractical as it requires the assistance of another person, has a limited number of words and phrases, and is very slow, which can be critical during an emergency situation.



(a)



(b)

Figure 1. (a): Patient using a communication board (b): Traditional Communication Board

To solve the aforementioned problem, we introduce an eye movement-based communication system for individuals with quadriplegia. The proposed system captures eye images and estimates the gaze vector, which represents where the person is looking. Using this predicted gaze vector, patients can efficiently communicate with medical assistants.

Gaze Estimation

Gaze estimation is a technology that determines where a person is looking by analyzing eye movements using cameras or infrared sensors and processing this data with advanced algorithms. By tracking eye features such as the pupil and corneal reflections, gaze estimation translates these movements into coordinates, allowing precise identification of the user's focus point. This technology has various applications, including enhancing human-computer interaction, enabling hands-free control for individuals with disabilities, improving user experience in virtual reality, assisting in psychological and cognitive research, and providing valuable insights for marketing and advertising by analyzing consumer attention.

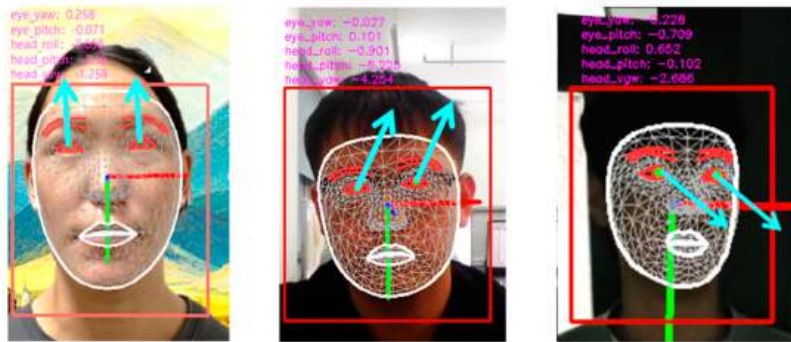


Figure 2. Example of gaze estimation and predicted gaze vector (Wei et al. 2023)

Communication Board

A communication board is a visual aid designed to facilitate communication for individuals with speech or motor impairments. It typically features a range of symbols, pictures, letters, or words that users can point to or select in order to convey their needs, thoughts, or feelings. The primary function of these boards is to provide a method of expression for individuals who cannot use verbal speech effectively, enabling them to interact with others and participate more fully in daily activities.

Quadriplegic patients, who experience partial or complete loss of use of all four limbs and the torso, face significant challenges in communication due to their impaired motor functions. Traditional communication methods, like speaking or gesturing, are often not viable. As a result, communication boards become a crucial tool for these individuals. By using eye movements, patients can indicate their choices on the board, allowing them to express needs, desires, or responses. However, this process can be laborious and time-consuming, as it often requires a caregiver to point at each symbol or image until the patient reacts, which can be particularly problematic in urgent situations.

To overcome the inefficiencies of traditional communication boards, gaze estimation technology offers an advanced solution. This technology tracks eye movements using parameters such as yaw and pitch to determine where a patient is looking. By integrating gaze estimation with a digital communication board, the system can quickly and accurately interpret the patient's gaze direction and match it with corresponding symbols or messages on the screen. For example, if a patient looks at an image of food, the system can detect this gaze and identify the patient's need as hunger. This method significantly enhances communication speed and efficiency which provides a more responsive and less cumbersome alternative to conventional boards, particularly benefiting those in urgent or critical situations.

Method

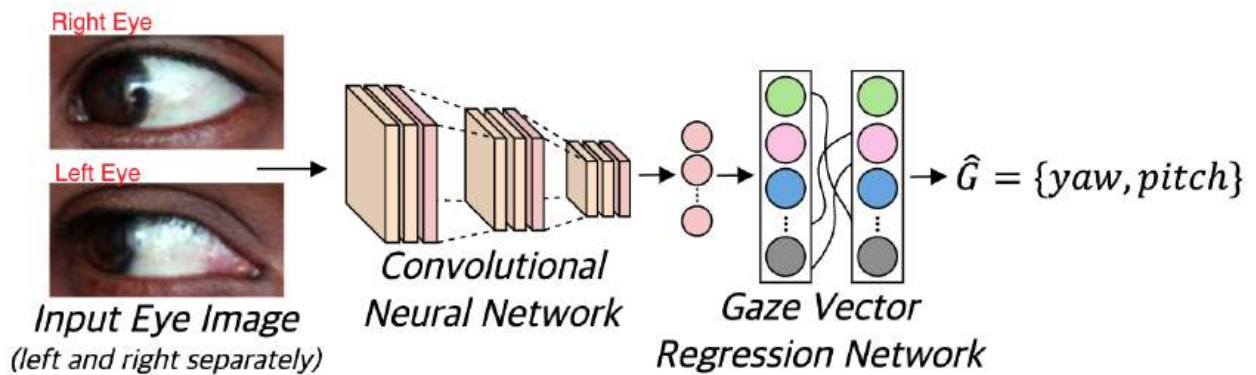


Figure 3. Architecture of the proposed gaze estimation network

Figure 3 illustrates the architecture of the proposed gaze estimation network. To train the module, the gaze vectors, yaw and pitch, are calculated after the image of the left and right eye goes through convolutional neural networks and regression networks. This value is set as the ground truth, **GT** for short, which later will be compared with the estimated value to calculate the loss. Getting the loss of the module depends on two factors, comparison with the ground truth and comparison between the left and right eye using conjugate gaze. Conjugate gaze is our eyes' ability to work in pairs. A person with no special medical condition has eyes that work in unison, therefore, our gaze vector in the left and right eye will be the same. Before comparing these factors, the module will compute the distance and angle of the gaze factors. In the formulas below, distance and angle is written as mean square error and degree.

Equation 1: Consistent mean squared error function

$$L_{Cons-MSE} = \frac{1}{N} \sum_{i=1}^N (\widehat{yaw}_i^L - \widehat{yaw}_i^R)^2 + (\widehat{pitch}_i^L - \widehat{pitch}_i^R)^2$$

To compute the mean squared error, MSE for short, the module will input the previous data, yaw and pitch, into the distance formula. After, the module will compare both sides of the eye to make sure conjugate gaze is functionated in the module.

Equation 2: Ground truth mean squared error function

$$L_{GT-MSE} = \frac{1}{2N} \sum_{k \in \{R,L\}} \sum_{i=1}^N (yaw_i - \widehat{yaw}_i^k)^2 + (pitch_i - \widehat{pitch}_i^k)^2$$

The module will compute the MSE in the same way as above, but compare the estimate with GT, which is the real value, to get the difference of the estimated distance from the GT.

Equation 3: Ground truth degree error function

$$L_{GT-Degree} = \frac{1}{2N} \sum_{k \in \{R,L\}} \sum_{i=1}^N \cos^{-1} \left(\frac{G_i \cdot \hat{G}_i^k}{|G_i| \times |\hat{G}_i^k|} \right)$$

To compute the degree, the module will input the yaw and pitch to the angle formula. After getting the calculated degree, the module will compare the estimate with the real value to get the difference.

Equation 4: Consistent degree error function

$$L_{Cons-Degree} = \frac{1}{N} \sum_{i=1}^N \cos^{-1} \left(\frac{\hat{G}_i^L \cdot \hat{G}_i^R}{|\hat{G}_i^L| \times |\hat{G}_i^R|} \right)$$

The module will use the same angle formula as above to get the degree, but compare the calculated values of the left and right eye to check the conjugate gaze.

Equation 5: Total loss function

$$L = L_{GT-MSE} + L_{GT-Degree} + \beta(L_{Cons-MSE} + L_{Cons-Degree})$$

Finally, the module will add all the equations it used above to calculate the loss. In this process, beta is represented as β and multiplied to the equations that compare the conjugate gaze because comparison between the ground truth and the module's estimate is far more important than calculating the conjugate gaze. In the module, beta is defined as 0.8 to reduce the influence conjugate gaze would have on the total loss.

Gaze Estimation Dataset

To train our proposed model we use the gaze capture dataset (Krafka et al. 2016), a large scale gaze estimation dataset that includes 2.5 million sample images. The images contain people of all races, gender, people who wear glasses and people who do not, in different terrain, lighting and pixel distribution. Figure 4 displays a sample of the gauge estimation model.



Figure 4. Example samples of Gaze Capture Dataset (Krafka et al. 2016)

Evaluation Metric

To evaluate the proposed approach, we utilize angular error function as explained in Equation 6.

Equation 6: Angular Error Function

$$Error = \cos^{-1}\left(\frac{G \cdot \hat{G}}{|G| \times |\hat{G}|}\right)$$

The Angular Error Function, compares the angles of the true gaze vector (G) and the estimation gaze vector from the proposed model ($\hat{G}$). The more accurate the estimation is, the difference in angle will be less, ideally the difference will be 0.

Results

In this chapter, we conduct three distinct experiments to evaluate the effectiveness of the proposed system. The first experiment involves a performance comparison with state-of-the-art gaze estimation methods (Figure 5). The second experiment is an ablation study, where we vary the beta values in Equation 5 to identify the optimal value that yields the best results (Figure 6). Lastly, the third experiment consists of user testing for the communication board. We implemented the digital communication board system using the proposed approach and assessed its accuracy across different setups (Figure 7).

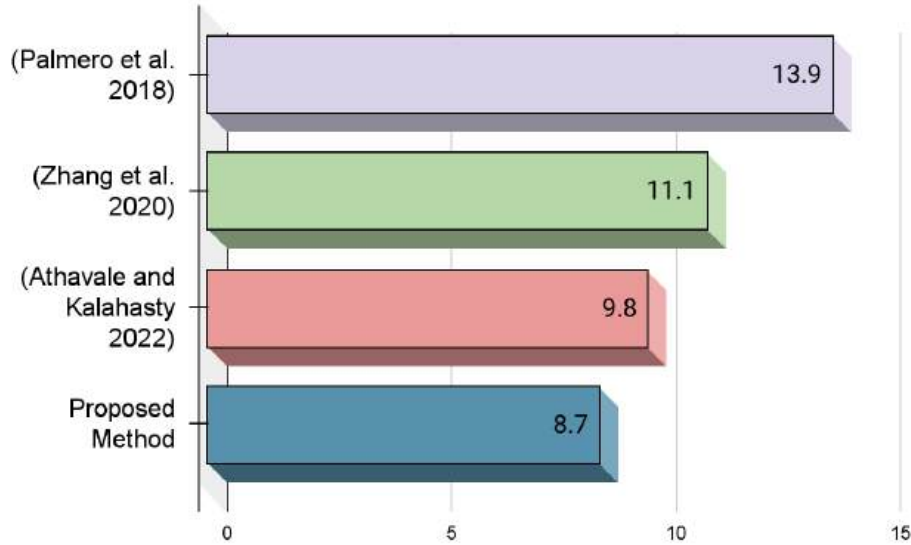


Figure 5. Performance comparison with state-of-the-art gaze estimation methods

Figure 5 illustrates the performance comparison between our proposed method and three state-of-the-art gaze estimation techniques. Among these, the proposed method achieved an error rate of 8.7, demonstrating superior accuracy compared to the most recent previous method, which had an error rate of 9.8.

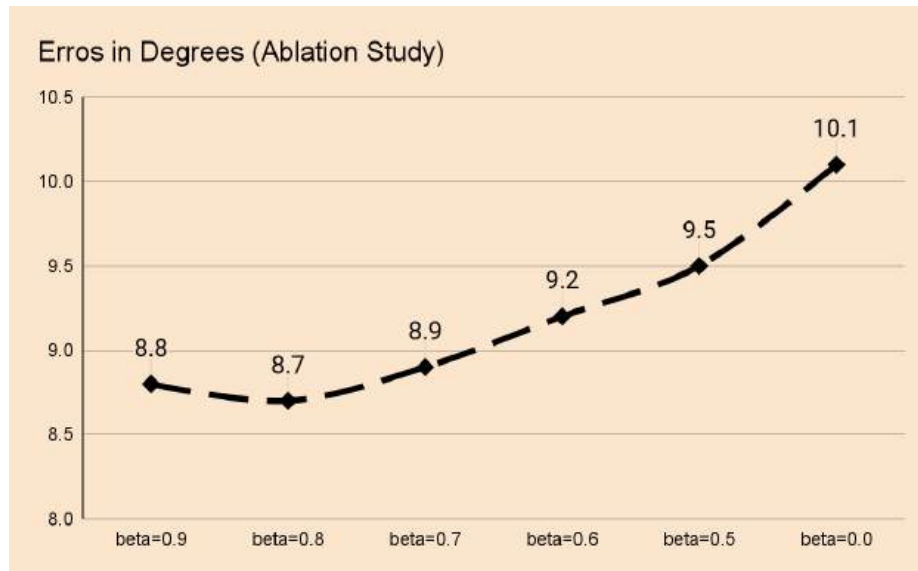


Figure 6. Analysis on the most optimal value for beta

To determine the optimal beta value in Equation 5, we conducted several experiments using different beta values. Figure 6 presents the evaluation results which indicates that setting beta to 0.8 yields the best performance. Small beta values produced poorer results, as they disrupted the model's ability to accurately predict the gaze vector. When beta was set to 0, conjugate gaze was not incorporated into the model, resulting in the highest loss, which was 10.1. This indicates that without the application of conjugate gaze, the model's accuracy is notably reduced. These experimental results clearly demonstrate the significance of conjugate gaze.

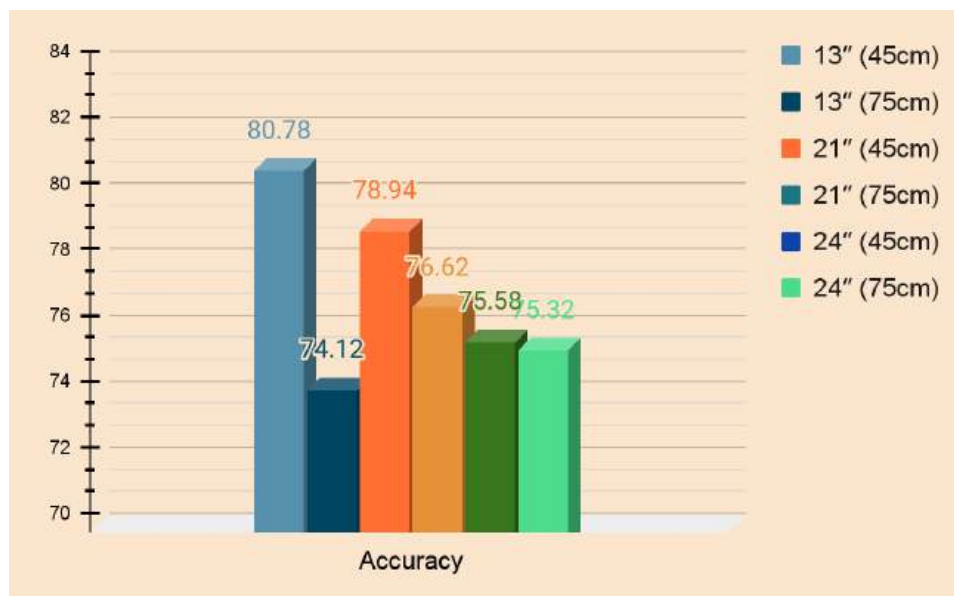


Figure 7. User test evaluation for different communication board setup

Finally, we conducted a user test to evaluate the efficiency and feasibility of the proposed approach in real-world scenarios. We implemented the approach across different digital monitor setups: 13-inch, 21-inch, and 24-inch monitors. Participants were instructed to select specific items on the communication board (monitor) from various distances, with the webcam positioned at the top of the monitor.

Figure 7 summarizes the experimental results. The 13-inch monitor, when used at a 45 cm distance, achieved the highest accuracy of 80.78%. However, accuracy decreased to 74.12% when the distance between the individual and the webcam was increased. For the 21-inch monitor, distances of 45 cm and 75 cm yielded accuracies of 78.94% and 76.62%, respectively. The 24-inch monitor showed similar accuracy across both distance setups.

Conclusion

In this study, we have developed and evaluated a novel gaze estimation-based communication system tailored for quadriplegic patients. By integrating a convolutional neural network with a gaze vector regression network, the proposed system minimizes prediction loss and accurately interprets eye movements. The experimental results indicate that the inclusion of conjugate gaze functionality significantly improves accuracy. The practical application of the proposed system was validated through user testing across various monitor sizes and distances. The 13-inch monitor setup demonstrated the highest accuracy at a 45 cm distance, though accuracy decreased with increased distance. The proposed approach also showed robust performance across different monitor sizes, confirming its adaptability to real-world conditions. Future work will focus on further optimizing the system, exploring additional use cases, and expanding its application to other assistive technologies.

Works Cited

- Athavale, R., Motati, L. S., & Kalahasty, R. (2022). One eye is all you need: Lightweight ensembles for gaze estimation with single encoders. ar
- Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., & Torralba, A. (2016). Eye tracking for everyone. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2176-2184).
- Palmero, C., Selva, J., Bagheri, M. A., & Escalera, S. Recurrent cnn for 3d gaze estimation using appearance and shape cues. arXiv 2018. arXiv preprint arXiv:1805.03064.
- Wei, J., Wu, H., Wu, Q., Iwahori, Y., Yu, X., & Wang, A. (2023). Gaze Estimation Method Combining Facial Feature Extractor with Pyramid Squeeze Attention Mechanism. Electronics, 12(14), 3104.
- Zhang, X., Park, S., Beeler, T., Bradley, D., Tang, S., & Hilliges, O. (2020). Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16 (pp. 365-381). Springer International Publishing.

Rethinking Spatial Design for Elderly Communities Across Contexts: The Case of Singapore and Montgomery County, Maryland, United States of America

By Aaron T. Chuang

Abstract

The growing trends of elderly populations and people living in cities formulates the issue of new needs in spaces that require design to adapt, emphasizing the importance of inclusivity and accessibility for marginalized communities in cities. This study will compare spaces in Montgomery County, Maryland, United States of America with Singapore and analyze ways in which spatial design addresses differences for the senior communities in both places. In addition, I will apply design principles from prominent authors in this subject, to each space analyzed through qualitative photographs taken by me, in five figures that each address a fundamental aspect of elderly life. Ultimately, while Maryland and Singapore are vastly different, the comparative analysis demonstrates that spatial design can be used to address differences through three key concepts, the outdoor vs. indoor dichotomy, the public vs. private space dichotomy, and the flexibility in use of a space.

Section 1: Introduction

The words, “senior” and “elderly” are often considered as classifications, and refer to not a biological stage, but a chronological age denoted as “old”. Seniors can be defined as an older person, usually over 60 years of age as stated by the United Nations (United Nations, 2025). However, seniors can also be defined as living historians; they have lived through a shift from community-centered and intimate architecture to denser and more technologized urban landscapes. According to the World Health Organization (WHO, 2024), it is expected that by 2050, the world’s population of people aged 60 and over will double from 1.4 billion to 2.1 billion. Not only is the world’s population of senior citizens increasing, but the number of people living in cities are also increasing.¹ This is in part due to new technologies in medical fields extending the typical lifespan, and new job opportunities. The escalating growth of seniors and residents of cities sets off an alarm, calling to attention that design does not adequately adapt to them, emphasizing the importance for inclusivity, and accessible cities. These needs include spaces that, for example, encourage physical activity, however with an emphasis on safety, and the prevention of falls or accidents.² These safety concerns regarding senior citizens in cities, particularly community spaces reminds the importance of addressing seniors through spaces and architecture.

Literature Review

¹ United Nations News, *New UN Report* (United Nations). The UN reported that about 2.5 billion more people will live in cities by 2050.

² U.S. Centers for Disease Control and Prevention: Older Adults Fall Prevention, About Older Adults Fall Prevention, (CDC, 2024). In 2021, falls caused 38,000 deaths among those aged 65 and above, and emergency departments reported over 3 million visits due to older adults falling.

Among all spatial design and architecture publications and research, three works from key thinkers, Victor Papanek, Ronald Mace, and Victor Regnier were selected. Victor Papanek was an Austrian-born American designer and educator, and one of the key figures in the advocacy for more socially and ecologically responsible design. Building upon Papanek's work, Ronald Mace, an American architect, product designer, educator, and consultant, formed principles guiding his term, "universal design". The last key thinker, Victor Regnier, American architect, professor, and researcher designed a set of guidelines drawing from both Papanek and Mace's principles for aging facilities. These works were selected because of the clear and direct frameworks they offer as a foundation, which directly inform and support the study on spatial design methods in designing senior communities. In addition, each study was published at a different chronological point, and are built upon each other, thus allowing for this study to be built upon each.

Victor Papanek, designer and educator published his book, "Design for the Real World" in early 1971, which introduced a new frontier, showing respect to social and environmentally responsible design. Papanek emphasizes the importance of serving human needs over commercial interests, arguing for design to prioritize physical, social, and cultural needs often neglected. He also promotes design for the marginalized and vulnerable, such as senior citizens. Furthermore, Papanek criticizes useless aesthetic designs, and encourages the highest priorities to be functionality, accessibility, and usefulness. Although he believed that most technology was wasteful, he thought that delicate use of appropriate technology would help spaces be more usable and influence personal choice among people of diverse backgrounds and abilities.³ Ultimately, Papanek opens a new concept of design, challenging designers to prioritize people over products, encouraging for design to be used as a tool to improve lives, shape society, and drive equity.

Ronald Mace, architect and product designer coined the concept of universal design which was later formalized through the "Seven Principles of Universal Design", published in the 1980s. Although there is no explicit evidence that Victor Papanek's theories and strategies influenced Mace's principles, both designers' works reflected a shared commitment for ethical, inclusive design as a response to similar cultural shifts in the 1970s. Mace's "Seven Principles of Universal Design" offers a direction from evaluating existing designs towards designing more usable products and environments to people of diverse abilities. All 7 principles are important and key in this research in design for senior citizens. In particular, principles 1, 2, 5, and 6 are important to all seniors.⁴ In contrast, principles 3, 4, and 7 directly influence citizens with

³ Papanek (1971) described most technology to be indulged in wasteful consumerism and focused only on serving the wealthy, he urged for the use of such "appropriate technology" that was more small-scale, local, and adaptable.

⁴ Principle 1 is equitable use, and emphasizes the need for all designs to provide the same means of use and appeal to all users. Principle 2, flexibility in use, explores the importance of flexibility in design that provides adaptability and independent choice amongst users. Principle 5, tolerance for error, guides for design to provide special features, arrangements/placements, and technologies, to ensure the minimization of hazards and errors. Principle 6, low physical effort, encourages design to ensure comfortability during use, allowing users to maintain a neutral body position.

physical, sensory, and cognitive challenges, highlighting the importance of inclusivity among different backgrounds and challenges.⁵ Overall, Mace pushes for the lesson that design should be usable by people of all abilities, and that physical spaces and products must ensure the user's accessibility, safety, and intuitive operation.

Finally, Victor Regnier, architect, professor, and researcher applied theories and principles from both Ronald Mace and Victor Papanek to the design for assisted living facilities, presenting key guidelines for housing the physically and mentally frail in his book, "Design for Assisted Living: Guidelines for Housing the Physically and Mentally Frail", published in 2002.⁶ Regnier favored home-like spaces with the main goal of placing older persons and their families first, before rules and regulations. Besides assisted living facilities, multiple of Regnier's principles can be applied to public community spaces. For instance, he believed that a single space should accommodate a variety of activities, so that residents could choose how to use them. In addition, Regnier stressed design to support physical accessibility, intuitive way-finding, and environmental comfort such as lighting, shade, and temperature to reduce barriers, ultimately serving the frail. Overall, Regnier emphasized that design should foster independence, dignity, well-being and safety, especially through access to nature, privacy, and social spaces.

Section 2: Methodology

The study aims to comparatively analyze two case studies, one in Singapore, and another in Montgomery County of Maryland, United States of America. Singapore and Maryland offer distinct yet similar contexts for studying senior-friendly spatial design. Although each case is shaped by its own cultural norms, climate conditions, and urban form, both provide unique opportunities for social engagement, mobility, and public life among older adults, and offer insights into how a community space designed for the same purpose is designed and adapted across two different settings.

The foundation of the research draws from the design theories of Victor Regnier, Victor Papanek, and the 7 Principles of Universal Design by Ronald Mace. The research data is collected through a qualitative collection of photographs of different locations/spaces in each case study. Data will be presented with photographs from each case study side by side, in 5 figures.⁷ These 5 figures were selected because each one addresses a fundamental aspect of

⁵ Principle 3, simple and intuitive use guides for the use of design to be easy to understand, despite diverse ranges of literacy and linguistic skills. Principle 4, perceptible information, guides for the design to effectively communicate its purpose to the user, and in particular utilize technologies and techniques for those challenged by limitations. Principle 7, size and space for approach and use, urges design to accommodate a variety of people by variations in body size, posture, and mobility.

⁶ Guidelines: 1. Appear residential in character, 2. Be perceived as small in size, 3. Provide residential privacy and completeness, 4. Recognize the uniqueness of each resident, 5. Foster independence/interdependence/individuality, 6. Focus on health maintenance/physical movement/mental stimulation, 7. Support family involvement, 8. Maintain connections with the surrounding community, 9. Serve the Frail

⁷ All photographs in the research are taken by Aaron Chuang.

elderly life, such as activities, gatherings, seating, movement, and privacy that is shaped differently based on geographical, climatical, and cultural differences.⁸

Case Studies: Singapore and Maryland, United States

The Republic of Singapore is a small city-state island nation in Southeast Asia near the equator. As of 2024, World Bank records Singapore's GDP to be US\$ 547.39 billion. Based on Singapore's Meteorological Service Singapore (World Bank, 2020), since it is near the equator, Singapore's climate is characterized by two monsoon seasons separated by inter-monsoonal periods, averaging about 31-33 degrees celsius throughout the year. In 2023, according to Data Commons, the nation was home to 5.918 million people (Data Commons, 2023), and in 2024, there were an estimated 754,000 senior residents, per Statistica (Statistics, 2024). The case study in Singapore, studies three main spaces. The spaces include: a hawker center/kopitiam, HDB community spaces which include an age-friendly playground or exercise station, and a shaded park with benches.⁹ A heavy focus is placed on hawker centers, particularly because they are protected by UNESCO, this reinforces that hawker centers play a dual role as both cultural heritage and as an everyday community space.¹⁰ This is important because it allows for discovery of the ways spatial design influences various factors for elderly communities. Overall, these three main spaces were chosen because they were commonly found within walking distance of an active aging center in Singapore and a block of condos, which formed a neighborhood of daily destinations for seniors. Their spatial relevance is reinforced by Singapore's hot climate, and mission to build accessible, shaded, and socially vibrant environments used to encourage outdoor activity and social engagement among seniors.

Maryland is a mid-atlantic state in the United States of America. According to the report of Data Commons, in Maryland, there are about 6.263 million residents (Data Commons, 2024), among which 1 million are senior citizens (Data Commons, 2023). In contrast to Singapore's climate, Maryland State Archive states that Maryland's climate is characterized by four seasons, spring, summer, autumn, and winter, with an average temperature of about 22.6 degrees celsius in the summer (Maryland State Archive, 2025). Instead of focusing on the entire state of

⁸ This study faces certain data limitations due to logistical and ethical constraints, such as time, respect of privacy, and limited access. Although the period of stay in Singapore was long enough for initial fieldwork, the time constraints and limited access to participants resulted in a smaller dataset than originally anticipated. Additionally, participation in sharing experiences and thoughts is voluntary, therefore it affects the representativeness and comprehensiveness of the findings. Therefore, the findings may not fully capture the diversity of experiences within the elderly community.

⁹Housing & Development Board, (*Singapore Government Agency, 2025*). The acronym "HDB" stands for the Housing & Development Board in Singapore, which provides public housing to Singaporeans . Its mission statement is "as Singapore's public housing authority, HDB builds homes and transforms towns to create a quality living environment for all". Typically, HDBs are located around hawker centers, commercial shops, shaded parks, places of gathering, and playgrounds for children as well as exercise stations for senior citizens.

¹⁰UNESCO Intangible Cultural Heritage, Hawker culture in Singapore, community dining and culinary practices in a multicultural urban context, (*UNESCO, 2020*). Hawker centers were community places "where people from diverse backgrounds gather and share the experience of dining over breakfast, lunch, and dinner" and play activities such as chess and busking.

Maryland, Montgomery County was chosen because its size more closely approximates to the size of Singapore, making it a better representative microcosm for the comparison. Additionally, Montgomery County's elderly population according to its government website, is expected to project to 14.4% (Montgomery County, 2020), about the same percent of Singapore's elderly population in 2024. The spaces studied were contrast to the ones in Singapore, and include: community recreation centers and outdoor shopping centers.¹¹ These spaces reflected different spatial models in which there is a larger car dependency, climate variation, and suburban layouts that influence how senior adults access public life and engage or interact with spaces and others.

Section 3: Comparative Analysis

Figure 1: Design for A Variety of Activities



Figure 1.1 compares a hawker center in Singapore and a recreation center in Maryland.

¹¹Montgomery County Recreation, Recreation Centers, (*Montgomery County Government, 2025*). Commonly within each county, there are multiple community recreational centers. "Montgomery County Recreation has 22 recreation centers...offer[ing] a wide range of activities to children, teens, and adults, and may include programs for seniors and individuals with disabilities.



Figure 1.2, Hawker Centers at Night and in the Afternoon

Figure 1.1, represents two spaces. First, an outdoor, fanned hawker center in Singapore, located about a 5 minute walk from an aging center, and about a 3 minute walk from a neighborhood of HDBs. Secondly, an indoor, air conditioned recreation center, about a 4 minute walk from a next door neighborhood, and about a 3 minute walk consisting of crossing a large road for the neighborhood in front. Even though many could walk there, almost all using the recreation center used a car to reach there. Figure 1.2 further shows the utility of the hawker spaces at night and in the afternoon.

Figure 1.1 demonstrates how spaces can accommodate multi-functional use enabling a range of physical, social, and cultural activities to take place in a single environment. Singapore's hawker centers typically comprise about 4 to 6 food stalls, selling a variety of cuisines from Chinese to Indian. In addition, the hawker center serves a purpose in which people from diverse backgrounds gather and play different activities such as chess, and busking. The hawker center promotes the importance of independent choice. For example, the hawker center is used for different activities. The table in the frontmost of the image in figure 1.1, shows a man slurping on his noodles, while the table next to it, a group of seniors are eating from small bowls, and chatting. Furthermore, in the images in figure 1.2, it is observed that some seniors are drinking beer, some are drinking coffee, while others are having a late night talk or dinner. Likewise, the recreation center in Maryland offers multiple activity rooms. The image shows one activity room with various activities from pool on the left, table tennis on the right, arcade games in the back, and extra seating for observation of the games. The recreation center promotes the use of independent choice, in which members of the recreation center are open to different types of activities depending on what they are interested in doing. These spaces embody the second principle of universal design, flexibility in use by providing the choice of different activities, and accommodating for user preferences, and group sizes. Whether the activities are gathering, eating, resting, or exercising, these spaces are adaptable to the changing needs of senior users throughout the day. In addition, Victor Regnier's principle of recognizing the uniqueness of each

person, is applied, since each person interacting with these spaces is unique by liking certain things. When a space is open and flexible to all, it recognizes the uniqueness of each person and promotes the importance of independent choice.

Figure 1.1 also highlights how intuitive design elements can enhance the usability for senior individuals, emphasizing that diverse users can navigate and interact with the space with ease. The hawker centers present menus shown in different languages and sizes, an addition to pictorial menus showing what each dish looks like. In his book, Victor Papanek enforced the importance of function over aesthetics, for example the presentation of the menus in the hawker center. Additionally, principle 3 and 4 can be implied to be applicable to the hawker centers and their menus. Similarly in Maryland, the recreation centers eliminate the unnecessary complexion of what each room is used for by having a large window looking into the room from the hallway. Although it may not look nice aesthetically, the principles of Victor Papanek, and Ronald Mace's are applied here, focusing on simple uses that emphasize functionality.

Though figure 1.1 has a lot of similarities of design elements, there is one clear difference. The hawker center in Singapore is open to all and relatively open early to late. On the other hand, Maryland's recreation center is only open to residents in Montgomery County, and for non-residents, they would have to pay an annual fee. In addition, the recreation center opened later and closed earlier than the hawker center. Among the take-aways of Victor Papanek, one stood out, the importance of designing to serve human needs over commercial interests. The hawker center stands as a place to serve human needs whether in early mornings or in late nights. However, the recreation center stands as an icon for only residents of certain communities, denying equitable use that is suggested as the first principle of universal design. In conclusion, the first pair represents spaces that are designed for a variety of uses, and each has design elements enforcing this idea.

Figure 2: Gatherings & Furniture



Figure 2.1 shows two gathering spaces: one directly below a condo with a table and benches, another with chairs in the center of a recreation center.

The gathering space below the condo features a mug in the foreground and a blue table with a built-in chess board in the background, suggesting a space that supports both casual conversation over drinks and more focused or quiet activities like chess. And in the recreation center, the movable chairs allow users to choose how they want to sit, gather, or engage with others. Figure 2.1 presents a demonstration of principle 2, flexibility in use, emphasizing the providing of choice in methods of use. Moreover, the chosen furniture emphasizes a focus on mental stimulation and access to social spaces, as promoted by the principles of Regnier. Additionally, the furniture's placement in the gathering-like space can be supported by Papanek's theory on design as a tool for social change, and his belief that spatial design can shape society, and improve lives through interactions.

The condo gathering space also shows materials to be made up of tiles which can be washed down easily, making it cleaner, and in terms of temperature, tiles reflect less heat compared to other materials, especially in warm climates such as Singapore. Similarly the recreation center displays the chairs to be made up of mostly leather, making it comfortable and soft. The choice of materials in the furniture promotes that design can be used efficiently and comfortably which is one of the guidelines of principle 6, low physical effort. The material choice can also be implied to be an application of Papanek's emphasis on sustainability and Regnier's belief of furniture and spatial design to reflect a home-like setting.

Finally, figure 2.1 can be perceived as accommodating for variations in body sizes, posture, and mobility. For example, the reddish-brown bench to the far right of the space in Singapore, accommodates both for members with mobility challenges such as wheelchair, and for those who want to sit on it. Adding on, the benches are curved which promote the tolerance for error. In parallel, the space in Maryland has chairs that are considerably wider than typical chairs, thus accommodating for variation in body size and posture. Mace's principle 7 (size and space for approach and use), and principle 5 (tolerance for error) are applied here, as well as Papanek's importance in designing for human needs, and for the recognition of uniqueness amongst all users via Regnier.

Figure 3: Idle Areas in Shopping Centers

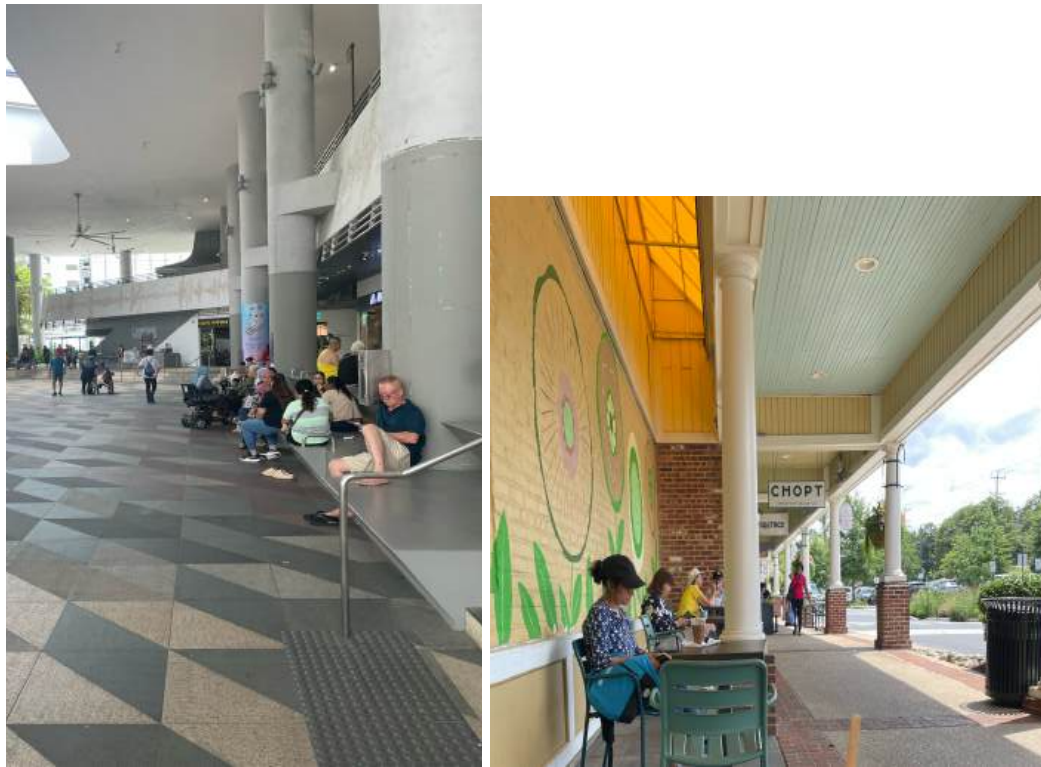


Figure 3.1 compares two idle areas in shopping centers: an open-shaded area on the ground floor of Kampung Admiralty, and a shaded area between shops at Wildwood Shopping Center in Maryland.¹²

Unlike figure 2 which were images taken of seating and gathering areas in residential amenity spaces, the third pair presents seating used as opportunities of rest in shopping centers. Figure 3 represents multipurpose idle areas in shopping centers. The idle area in Singapore displays seating in tiered long benches with handles. The benches form a circular outline around the center of the shopping area. Similarly, in Maryland, the space displays a seating area arranged with tables and chairs placed under shade in a strip mall. Figure 3.1 showcases the seating being used by a diversity of people, throughout age and challenges to ability. In addition, both locations are open to the public. In Singapore, the seating areas are made up of similar materials and color schemes throughout, including tiles, concrete, and metal. In Maryland, each table is the same size, and has around the same number of green metal chairs.

Both spaces also do not have a particular use, and showcases different uses for each. The area in Singapore displays that various activities can occur on each of the large benches and in the space entirely with the different shops surrounding, combining with the large space in the center in which seniors with mobility issues could roam around easily. For example, in the far center of the space in Singapore, it is observable to see a man in a wheelchair being pushed around the open space, and a woman in an electric power wheelchair chatting with her friends

¹² "Singapore's first integrated vertical elderly and co-living facilities for residents of all ages"

who are seated on the bench. Maryland's idle area also shows the flexible uses of the tables. For example, the lady nearest is seen reading a book, while the lady in the middle is on her phone, and the senior couple in the back are eating. Additionally, since the chairs are movable, a person with mobility issues will likely be able to sit alongside on these tables, proving the space's size and space for approach and use.

Furthermore, the idle area in Singapore shows the benches to be relatively close to the ground, limiting errors of falling. But, since the benches are tiered, there are risks of falling backwards. In addition, there are tactile pavings on the floor as perceptible information for those limited by eyesight, and other challenges. The image presents many resting and shows handles on benches to allow for comfortable movements from sitting to getting up and vice versa. Similarly in Maryland, although there are no tactile pavings, there are changes in material. However, in contrast, the shopping is built one step higher than the ground, creating a higher risk than the idle area in Singapore in terms of errors particularly in falling or tripping.

These spaces offer an application of a variety of principles and theories. For instance, Ronald Mace's principles 1, 2, 4, 6, and 7 are combined and demonstrated in figure 3.1. Based on such principles, a well-designed space is one that treats all users fairly, adapts to varied needs, communicates information clearly, minimizes physical strain, and offers ample room to move, creating an environment where everyone can participate and feel at ease. Additionally, the approach aligns with Papanek's call for socially responsible, and human-centered design that responds to the needs of people instead of aesthetics and trends. Further, Regnier's emphasis on environments that support the physical and cognitive, as well as emotional wellbeing of the ageing populations through thoughtful spatial and material choices are displayed in both spaces.

However the spaces are not as thoughtful in the tolerance for error. Though the spaces provide handles, tactile pavings, some parts of the space are more likely to cause error. For example, the tiered benches in Singapore and the uneven grounding of the strip mall in Maryland, can cause falling or tripping issues. This contradicts the universal design aim to minimize hazards and does not meet Papanek and Regnier's expectations of a space that prioritizes safety for all users, especially the frail and those more vulnerable to injury.

Figure 4: Spaces to Enhance Physical Activity



Figure 4.1 presents an exercise facility in the center of a pavilion, park, and apartments in Singapore, and a hallway inside a recreation center in Maryland connecting a mainstream walkway to various activity rooms.

The spaces presented in figure 4.1 utilize design principles and elements that encourage physical activities such as walking, and using exercising machines by providing accessible, safe, and engaging environments that support movement and exercise to support the physical well-being of the seniors. The exercise station is open to the public, and can be used by seniors as well as other ages. Additionally, the environment of the space is appealing to multiple types of users, from seniors exercising in the stations, to people walking around the pavilion, and to children playing in the playground. Although the recreation center in Maryland is stated to be only for members of the recreation center, any of those members are open to using these hallways for any uses. It can be implied that since the chairs are placed next to the wall, that these spaces provide a clear line of sight for important elements, as well as adequate space for the use of assistive devices or personal assistants.

Similarly, the environment in the park also provides enough space, thus a woman is in a wheelchair and interacting with the exercise station. Principles 1 (equitable use), 2 (flexibility in use), and 7 (size and space for approach and use) can be applied to figure 4 because both spaces are usable by many, and are flexible in the type of activities that occur with the space. In addition, the spaces are all sized adequately for easy and accessible approach and use. Many times, Victor Papanek emphasized the importance of functionality in designs and for designs to be tools to improve lives. The way these spaces were designed builds upon Papanek's fundamentals, these spaces can be used and functioned in various ways, and improve the lives of seniors by promoting and encouraging physical activities.

One main difference between the spaces is that the hallway presents a space that allows for safe use and pictorial instructions of what each activity room connecting to the hallway is used for. For example, on the right side of the hallway, it can be implied that the room is a large studio that can be used as a gymnasium or a dance studio because of the window that offers a

pictorial instruction. Additionally, the placement of the chairs can be understood as to avoid users walking through, hurting and injuring themselves, allowing for direct walking. On the other hand, the exercise station in the park of Singapore, does not offer any pictorial, verbal, or tactile instruction on how to use each equipment. And, the equipment does not provide safety measures to ensure errors do not occur. The spaces, since promoting physical activities, also utilize Regnier's principle of supporting physical movement to ultimately serve the frail. In addition, principle 4 (perceptible information), and principle 5 (tolerance for error) enforce these rules.

Figure 5: Seating in the Public vs. Private Dichotomy



Figure 5.1 presents two seating spaces in a public vs. private dichotomy, one represents a shaded pavilion utilized as a place of gathering and rest, the other represents a house in Maryland with chairs on its shaded front porch.¹³

Similar to the other community spaces explored in this study, a group of four seniors are seen together, chatting. The design of the space is equitable to all because of the similar materials used throughout the furniture in the space. Furthermore, the pavilion displays an implication that the spatial design is appealing to all users of various age ranges, since a stroller is in the far back. It is also observable that two of the seniors are sitting in plastic yellow chairs, while the other two are on metal benches. One senior has placed their feet up in a laying position, while another is sitting with their legs crossed. Finally the space presented shows even grounds, limiting risks of falls and promoting safety, a principle that Ronald Mace presented, principle 5, a tolerance for errors.

Although it can be seen that three principles of Mace, principle 1: equitable use, principle 2: flexibility in use, and 5: tolerance for error can all be applied to the space in Singapore, the space in Maryland contrasts and does not offer an example of the principles. Firstly, the house is a private residence, thus people who do not live in the home, can not simply sit on the residences' porch. Therefore, the space is not equitable or open to all. Secondly, the space is not flexible in use. Since the porch is relatively thin, the chairs would not necessarily be easy to

¹³ Figure 5.1 has a limitation in which it is only implied that people utilize the chairs on the front porch of the house for idling, and gathering.

move around, thus users are not able to customize the seating to face each other. Whereas in the pavilion, users can move around the chairs however they like. Thirdly, the porch is elevated from the ground and does not have any handles nor rails, so there is a higher risk of error, in particular, falls. Therefore, the design of the spaces in this pair contrast, and present different examples of the application of certain principles. The design of the spaces also explored the public vs. private dichotomy.

Section 4: Discussion

Culturally, Singapore and Maryland are vastly different, yet similar. However, the agency of design is not fundamentally concentrated on cultural lifestyle differences between Singapore and Maryland as forecasted by the research question, but by geographical and climatical differences. Singapore is a small nation compared to Maryland and the United States, thus many necessities and family members live in close proximity to seniors (World Data, 2025).¹⁴ Additionally in Maryland, many of the neighborhoods are about a 10 minute drive from parks, and a 30 minute drive for groceries or shopping. This contrasts heavily from Singapore's geographical size. Thus, the research question is partly incorrect. Rather than finding ways spatial design addresses cultural differences, the research shows the design of spaces focuses, and adjusts to geographical differences, as it is more important.

Section 5: Conclusion

Drawing from the comparative analysis, the three main ways in which spatial design addresses differences for the senior communities in Singapore and Maryland are, 1) The outdoor and indoor dichotomy, 2) The public and private space dichotomy, and 3) Flexibility in space. Firstly, in the figures, most seniors were found gathering, and interacting with outdoor spaces over indoor ones in Singapore. On the contrary, most seniors in Maryland were found interacting and exercising in indoor spaces. Secondly, many of the spaces studied in Singapore were public community spaces in which anyone could have access to, via car, walking, or train. While in Maryland, most of the spaces such as the activity rooms in the recreation center were only open for residents or special members. Finally, both spaces in Singapore and Maryland emphasized independent choice through flexibility in uses. For example, Singapore's hawker centers provided various cuisines, while the recreation centers provided various activities for members to participate in. These design strategies are not just subjective to seniors but can be implemented into non-senior contexts. For instance, the recreation centers in Maryland did not just have seniors interacting, but children and adults interacted with the seniors in playing pool and table tennis. Furthermore, families meeting with grandparents are normally found gathering in hawker centers in Singapore. By observing the ways people of different age groups interact with space, it is possible to imply that these three design strategies are open for all age-groups and peoples of different abilities.

¹⁴WorldData.Info, Country Comparison, (*WorldData.Info*, 2025). Singapore is about 735.7 sq km according to the Singapore Statistics Department (2024), while Maryland is a state in the United States, a nation approximately 9,525,067 sq km.

Based on the limitations, there are further questions open to study. In the future, through interviews and surveys with the senior population, researchers can further understand how spatial design addresses differences in the senior communities. How can spatial design elements influence patterns among senior individuals across different cultural and climatic contexts? In what ways does spatial design mediate the social life of seniors in community spaces around Singapore and Maryland? Differences in how seniors utilize the community spaces in Singapore and Maryland reveal how spatial design mediates aging, pushing towards further studies in the role of spatial design and architecture in social inclusion.

Works Cited

- Emergency Handbook Team. "Older Persons" [Older Persons]. UNHCR: The UN Refugee Agency. UNHCR. Last modified June 13, 2025. Accessed August 12, 2025.
<https://emergency.unhcr.org/protection/persons-risk/older-persons>.
- Google. "Maryland" [Maryland]. Data Commons. Accessed August 12, 2025.
https://datacommons.org/place/geoId/24?utm_medium=explore&mprop=count&popt=Person&hl=en.
- Google. "Singapore" [Singapore]. Data Commons. Accessed August 12, 2025.
https://datacommons.org/place/country/SGP?utm_medium=explore&mprop=count&popt=Person&hl=en.
- Government of Singapore. "Environment" [Environment]. Department of Statistics Singapore. Last modified 2025. Accessed August 12, 2025.
<https://www.singstat.gov.sg/publications/reference/ebook/society/environment#:~:text=The%20total%20land%20area%20of,Water%20Mark%20cadastral%20survey%20boundaries>.
- Mace, Ronald, and North Carolina State University. "The 7 Principles" [The 7 Principles]. Centre for Excellence in Universal Design. CEUD. Last modified 1980s. Accessed August 12, 2025. <https://universaldesign.ie/about-universal-design/the-7-principles>.
- Maroquin, Nelson. "5 Types of Chair Upholstery Fabric" [5 Types of Chair Upholstery Fabric]. World Linen Fabrics. World Linen. Last modified June 21, 2021. Accessed August 12, 2025.
https://worldlinen.com/blogs/news/5-types-of-chair-upholstery-fabric?srsId=AfmBOopvD6oZUPbMU1JuWjcVF_IPq8qJte3gFDnty3F5DryEfef_wSg.
- Maryland Government. "Maryland at a Glance" [Maryland at a Glance]. Maryland Manual On-line: A Guide to Maryland and Its Government. Last modified February 3, 2025. Accessed August 12, 2025.
<https://msa.maryland.gov/msa/mdmanual/01glance/html/weather.html>.
- Montgomery County Government. "Chapter II Demographics" [Chapter II Demographics]. Montgomery County Government. Accessed August 19, 2025.
https://www.montgomerycountymd.gov/senior/Resources/Files/16_senior_demographics_chapterii.pdf.
- Montgomery County Government. "Recreation Centers" [Recreation Centers]. Montgomery County Recreation. Last modified 2025. Accessed August 12, 2025.
<https://www.montgomerycountymd.gov/rec/facilities/recreationcenters/>.
- Papanek, Victor J. *Design for the Real World : Human Ecology and Social Change*. Academy Chicago Publishers, 1971.
- Regnier, Victor. *Design for Assisted Living : Guidelines for Housing the Physically and Mentally Frail*. J. Wiley, 2002.
- Singapore Government. "Climate of Singapore" [Climate of Singapore]. Meteorological Service Singapore. Accessed August 12, 2025.
<https://www.weather.gov.sg/climate-climate-of-singapore/#:~:text=The%20daily%20temperature%20range%20has,hour%20mean%20of%2026.8%C2%BAC>.

- Singapore Government. "Housing and Development Board" [Housing and Development Board]. Housing and Development Board. Last modified 2025. Accessed August 12, 2025. <https://www.hdb.gov.sg/cs/infoweb/homepage>.
- UNESCO. "Hawker Culture in Singapore, Community Dining and Culinary Practices in a Multicultural Urban Context" [Hawker Culture in Singapore, Community Dining and Culinary Practices in a Multicultural Urban Context]. UNESCO Intangible Cultural Heritage. Last modified 2020. Accessed August 12, 2025. <https://ich.unesco.org/en/RL/hawker-culture-in-singapore-community-dining-and-culinary-practices-in-a-multicultural-urban-context-01568>.
- UN News. "Around 2.5 Billion More People Will Be Living in Cities by 2050, Projects New UN Report" [Around 2.5 Billion More People Will Be Living in Cities by 2050, Projects New UN Report]. United Nations. Accessed August 12, 2025. <https://www.un.org/en/desa/around-25-billion-more-people-will-be-living-cities-2050-projects-new-un-report#:~:text=Calendar-,Around%202.5%20billion%20more%20people%20will%20be%20living%20in%20cities,urban%20planning%20and%20public%20services>.
- USA.Gov. "About Older Adult Fall Prevention" [About Older Adult Fall Prevention]. U.S. Centers for Disease Control and Prevention. Last modified May 16, 2024. Accessed August 12, 2025. <https://www.cdc.gov/falls/about/index.html#:~:text=Falls%20among%20adults%2065%20and,injury%20death%20for%20that%20group.&text=In%202021%2C%20emergency%20departments%20recorded,visits%20for%20older%20adult%20falls>.
- World Bank Group. "GDP (Current US\$) - Singapore" [GDP (Current US\$) - Singapore]. World Bank Group Data. World Bank Group. Last modified 2025. Accessed August 12, 2025. <https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?locations=SG>.
- World Health Organization. "Ageing and Health" [Ageing and Health]. World Health Organization. WHO. Last modified October 1, 2024. Accessed August 12, 2025. <https://www.who.int/news-room/fact-sheets/detail/ageing-and-health>.

Predicting Sentiment Score in Social Media Trends Using Artificial Intelligence

By Rishma Desai

Abstract

Understanding what impacts how content is perceived has been a crucial investigation for researchers as well as marketing professionals in the business world. In this study we explored whether machine learning could be a reliable solution. We tested four different models, a Decision Tree classifier, Random Forest classifier, Multi-Layer Perceptron classifier, and a K Nearest Neighbors classifier, to assess their efficacy in predicting the sentiment a post received based on engagement metrics and post characteristics. The results demonstrated that the models, specifically the Decision Tree and Random Forest Classifiers, were able to predict these scores with strong accuracy, precision, and recall. Our research proves that machine learning models have significant potential to enhance brands' ability to improve audience engagement through data-driven insights.

Introduction

Recently, the use of social media in digital advertising has seen rapid growth. Nearly 64% of the global population uses social media, creating an ideal platform with widespread reach for brands to make themselves known. As businesses increasingly utilize platforms such as Facebook, Instagram, and X (Twitter) to engage with consumers and develop their image, global spending on social media advertisements has skyrocketed, and was expected to reach \$270 billion by 2025 ("Social Media Advertising and Marketing Worldwide" 2023). Social media's popularity as a marketing device stems from its usage of key features such as likes, comments, hashtags, and shares, which are common to most major platforms. These interactive elements allow users to engage directly with content posted on the platform, and provide immediate feedback to the content creator. This input is critical for marketers because audience sentiment plays a central role in shaping brand perception, influencing engagement, and ultimately determining the efficacy of content (Singh 2022).

While social media marketing as a general field has been well-researched, information regarding the relationship between sentiment and engagement outcomes is still limited. This creates an opportunity to explore data-driven approaches, specifically those involving artificial intelligence (AI) and machine learning (ML) to research and predict audience sentiment as a step toward optimizing campaigns for reach and impact.

Previous studies have shown significant potential for AI models in predicting content performance on social media. The study *Unleashing Virality: A Data-Driven Exploration of Video Content Success through AI and ML Integration on Social Media Platforms* (2023) explores how different factors such as engagement metrics, post timing, and content features influence a model's ability to predict outcomes. This paper examined the concepts behind using ML models for this purpose, but did not develop or test an actual model. A study by Rajapaksha and Silva (2023) evaluated several models and their capacity to correctly predict virality on

Instagram posts. Their findings indicated the Random Forest Regression model performed the best, achieving an accuracy score of approximately 85%. Although research in this area exists, much of it has focused narrowly on predicting virality or engagement metrics, with relatively less emphasis on sentiment prediction despite its importance in understanding how content resonates with audiences.

The purpose of this project is to build on these studies by applying machine learning methods to predict the sentiment of social media content using a publicly available dataset. Specifically, the aim is to evaluate whether models such as decision trees, random forests, multilayer perceptrons, and K-nearest neighbors, as well as a mixture of experts model, can reliably predict viral potential based on post features. By integrating multiple models and evaluation metrics, this study will test the hypothesis that by combining content and engagement-based features, a machine learning model will be able to fairly accurately predict the sentiment of a post. On a personal level, this project serves as a valuable opportunity to connect my interest in the business marketing sphere with AI's transformative potential. By bridging these two fields, this paper aims to highlight how machine learning tools can be utilized to assist real-world marketing decisions and increase marketing efficiency.

Materials and Methods

For this study we utilized a publicly available dataset from Kaggle, found [here](#), as the basis of our research. It was composed of post samples from Twitter, Instagram, and Facebook, with time-related metrics like *post date* and *post time*, as well as engagement metrics such as *likes* and *comments*. Additionally, each data entry was assigned a sentiment score of *positive*, *negative*, or *neutral*. Although this dataset served as a solid foundation for analysis, its limited size made data augmentation necessary to control the risk of overfitting due to lack of sufficient data. We then split the data into 75% training data and 25% testing data to ensure there is sufficient data for the models to pick up patterns while still setting a quantity aside to evaluate their performances.

We compared the performance of four different models: Decision Tree classifier, Random Forest classifier, K Nearest Neighbors classifier, and Multi-Layer Perceptron classifier.

The Decision Tree classifier splits the data into increasingly smaller subsets based on the most significant features in the data. Each internal node tests an attribute, each branch represents the outcome of the test, and each leaf node provides a final classification, in this case whether the sentiment is *positive*, *neutral*, or *negative*. The model's *max depth* parameter can be altered to maximize output and obtain clearer classifications.

The Random Forest classifier combines multiple Decision Tree models each trained on a random subset of data into an ensemble and aggregates the classifications of the Decision Trees into a majority vote. Due to the variety of the models, it typically has higher accuracy than any single Decision Tree classifier. The parameters used to modify this model are max depth and number of estimators.

Rather than learning explicit patterns during training, the K-Nearest-Neighbors model classifies a new data point by identifying the closest data points from the training set, and assigning it the majority classification among those neighbors. The primary parameter for this model is K , the value that indicates how many of the closest data points the model should consider.

A Multi-layer Perceptron is a type of neural network composed of multiple layers of neurons. Each neuron passes information to every neuron in the next layer through a weighted connection, and the model learns by adjusting these weightings to minimize error between the predictions and actual classifications. The parameter we chose to define for this model was max iterations, which limits the number of times a calculation can be repeated.

Overfitting occurs when a model performs significantly better on the training data than on testing data. This means the model has memorized the training set rather than learning patterns it can use to predict. A small gap between training and testing accuracy is acceptable and often expected; but if the training accuracy is significantly higher than testing accuracy, it strongly indicates overfitting. The more complex a model is the higher the likelihood of overfitting, because it sets parameters based on specific details of the training data. For this reason, the goal is to choose the most complex model possible with the least disparity in training and testing accuracies.

A commonly used technique to counteract overfitting is called regularization. This technique, which was applied to the MLP model, penalizes models that rely too heavily on large weights, which overinflates the training accuracy. In L1 regularization, the sum of the weights is subtracted from the model's accuracy. This discourages models from assigning extreme importance to a small number of features. For example, if a model achieves 95% accuracy but the sum of the weights is 5, the penalty lowers its effective accuracy to 90%. Similarly, a model with 98% accuracy but a sum of weights of 10 would see its score drop to 88%. Although the second model achieves a higher accuracy, after the weight penalty is applied it performs worse than the first model with a lower initial accuracy. L2 regularization applies the same principle in a different way. It encourages a model's weights to be distributed more evenly across features, ensuring that if one feature is missing or less informative, the model can still make accurate predictions using the remaining features. It does this by squaring the individual weights before subtracting them from the accuracy. If one feature has a weight of 10, the penalty becomes 100, substantially reducing the model's accuracy. On the other hand, if the same total weight is spread evenly across multiple features, the penalty remains smaller, promoting more balanced feature importance.

After evaluating multiple models, we investigated whether a combination has a better performance than the individual models. In our research we took a simple average, however, the result can be skewed by outliers, and tends to reflect a compromise rather than the strongest prediction.

Alternatively, by taking a weighted average, the models that perform better will be represented more and the models that perform poorly will have less influence. In a classification

setting, imagine training ten models, six of which predict correctly on a test case. For the next round, whenever a model makes an error, its weight is reduced by half. Over time, models that consistently misclassify lose influence, while those that reliably predict correctly retain or gain greater weight. This adaptive reweighting strategy makes the ensemble more robust than simple averaging, as it directly penalizes errors and rewards accuracy. In a regression case, the same principle applies but with a range of error outputs rather than binary predictions. Instead of halving the weight each time a model misclassifies, the weight is adjusted based on the magnitude of its error. For example, when predicting temperature if the model is off by 3 degrees, the weight is divided by $2^3=8$. The smaller the margin of error, the less the model will be penalized, and vice versa. This way, models that are more precise retain greater weight in the final prediction, leading to more accurate overall results. The efficacy of this adaptive reweighting approach can be seen as an open question for future exploration.

These models were then evaluated on a series of metrics to find the most comprehensive reading of their abilities. The first metric, *accuracy*, determines the total correct predictions amongst all of the predictions. The equation for this is:

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

Accuracy can be misleading in datasets with class imbalance (e.g., 95% of posts are neutral), as a model that simply predicts the majority class all the time would still achieve a high accuracy. Therefore it is prudent to use other metrics to determine the efficacy of the models. The next metric we implemented was *precision*, which determines how many of the positive predictions were actually positive.

$$precision = \frac{TP}{TP + FP}$$

High precision indicates that the model is accurate when it identifies positive instances, making this metric especially important when false positives are costly or undesirable. Another metric we employed was *recall*. Also known as sensitivity, recall measures the percent of true positives the model correctly identified.

$$recall = \frac{TP}{TP + FN}$$

This is most beneficial in cases where missed positive cases (or negative/neutral depending on the focus) are consequential. Finally, the last metric we applied was *F1 score* which combines precision and recall into one metric.

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

This score is particularly useful when dealing with class imbalance, or if there is a need to balance precision and recall. Altogether, these assessment metrics provide a comprehensive understanding of a model's performance, and help target specific areas of improvement for further testing and development.

Results

Decision Tree with Max Depth = 23

	precision	recall	f1-score
0	0.98	0.95	0.97
1	0.94	0.91	0.93
2	0.96	0.99	0.97
	accuracy		0.96
macro avg	0.96	0.95	0.96
weighted avg	0.96	0.96	0.96
	precision	recall	f1-score
0	1.00	1.00	1.00
1	1.00	1.00	1.00
2	1.00	1.00	1.00
	accuracy		1.00
macro avg	1.00	1.00	1.00
weighted avg	1.00	1.00	1.00

This model is overfitting which can be understood by the perfect scores on all of the training data. This shows us that the optimal max depth parameter should be less than 23.

Decision Tree with Max Depth = 7

	precision	recall	f1-score
0	0.98	0.95	0.97

	1	0.92	0.94	0.93
	2	0.96	0.98	0.97
accuracy				0.96
	macro avg	0.96	0.95	0.95
	weighted avg	0.96	0.96	0.96
precision recall f1-score				
	0	1.00	0.99	0.99
	1	0.96	0.99	0.98
	2	0.99	0.98	0.99
accuracy				0.99
	macro avg	0.98	0.99	0.99
	weighted avg	0.99	0.99	0.99

Random Forest with Max Depth = 9, and Number of Estimators = 9

precision recall f1-score				
	0	0.92	0.95	0.94
	1	0.96	0.80	0.87
	2	0.94	0.99	0.97
accuracy				0.94
	macro avg	0.94	0.92	0.92
	weighted avg	0.94	0.94	0.94
precision recall f1-score				
	0	0.98	0.98	0.98
	1	0.98	0.93	0.96
	2	0.97	1.00	0.98
accuracy				0.97
	macro avg	0.98	0.97	0.97
	weighted avg	0.98	0.97	0.97

Random Forest with Max Depth = 7, and Number of Estimators = 3

	precision	recall	f1-score
0	0.81	0.82	0.82
1	0.70	0.59	0.64
2	0.88	0.94	0.91
accuracy			0.82
macro avg	0.80	0.78	0.79
weighted avg	0.82	0.82	0.82

	precision	recall	f1-score
0	0.84	0.85	0.85
1	0.78	0.73	0.76
2	0.90	0.92	0.91
accuracy			0.85
macro avg	0.84	0.84	0.84
weighted avg	0.85	0.85	0.85

MLP with Max Iterations = 300

	precision	recall	f1-score
0	0.86	0.59	0.70
1	0.35	0.33	0.34
2	0.69	0.85	0.76
accuracy			0.65
macro avg	0.64	0.59	0.60
weighted avg	0.66	0.65	0.65

	precision	recall	f1-score
0	0.83	0.59	0.69

1	0.44	0.39	0.41
2	0.69	0.87	0.77

accuracy			0.66
macro avg	0.66	0.62	0.63
weighted avg	0.66	0.66	0.65

MLP with Max Iterations = 100

	precision	recall	f1-score

0	1.00	0.03	0.05
1	0.08	0.01	0.02
2	0.50	1.00	0.66

accuracy			0.48
macro avg	0.52	0.35	0.25
weighted avg	0.55	0.48	0.33

	precision	recall	f1-score

0	0.75	0.02	0.04
1	0.27	0.04	0.06
2	0.47	0.99	0.63

accuracy			0.46
macro avg	0.50	0.35	0.24
weighted avg	0.49	0.46	0.31

K Nearest Neighbors with Number of Neighbors = 5

	precision	recall	f1-score

0	0.56	0.65	0.60
1	0.38	0.37	0.38
2	0.74	0.68	0.71

accuracy			0.60
macro avg	0.56	0.57	0.56
weighted avg	0.60	0.60	0.60

	precision	recall	f1-score
0	0.67	0.78	0.72
1	0.65	0.58	0.62
2	0.81	0.78	0.80

accuracy			0.73
macro avg	0.71	0.72	0.71
weighted avg	0.73	0.73	0.73

K Nearest Neighbors with Number of Neighbors = 22

	precision	recall	f1-score
0	0.60	0.63	0.61
1	0.45	0.38	0.41
2	0.76	0.79	0.77

accuracy			0.65
macro avg	0.60	0.60	0.60
weighted avg	0.64	0.65	0.64

	precision	recall	f1-score
0	0.61	0.66	0.63
1	0.54	0.41	0.47
2	0.71	0.79	0.75

accuracy			0.65
macro avg	0.62	0.62	0.62
weighted avg	0.64	0.65	0.64

When we aggregate these models by taking an average score, the combined models have a 75% accuracy.

Discussion

This paper proved machine learning models are accurately able to predict sentiment in posts on Instagram, Facebook, and Twitter based on their engagement and post detail features.

The Decision Tree and Random Forest models performed better than the others, achieving a 96% accuracy.

To further this academic research, we would like to investigate whether the accuracy of the average model can be improved using a weighted system when further developing this model. We would also like to explore expanding the model to encompass posts from other platforms. It currently does not include post data from TikTok, which is a prominent platform for brand marketing. Additionally, future work could compare and contrast trends across various countries to see how content fares in different cultural settings. Likewise, it would be beneficial to explore real world applications of this model in the marketing field, possibly by testing the models' abilities to predict sentiment without engagement metrics, relying purely on controllable variables like post time and day.

Works Cited

- Jehnsin, Ben. Survey of Best Regularization Techniques in Artificial Intelligence: From Classical Methods to Modern Deep Learning Approaches. 3 Jan. 2025, www.researchgate.net/publication/394255283_Survey_of_best_Regularization_Techniques_in_Artificial_Intelligence_From_Classical_Methods_to_Modern_Deep_Learning_Approaches.
- Lakshmi, Namratha, and Vempaty. Unleashing Virality: A Data-Driven Exploration of Video Content Success through AI and ML Integration on Social Media Platforms.
- Rasika Rajapaksha, and Singhe Silva. "Predictive Analysis on Social Media Content to Become Viral." International Conference on Advances in Computing and Technology 2024, 15 Dec. 2023, www.researchgate.net/publication/381852269_Predictive_Analysis_on_Social_Media_Content_to_Become_Viral. Accessed 11 Aug. 2025.
- scikit-learn. "Sklearn.model_selection.train_test_split — Scikit-Learn 0.20.3 Documentation." Scikit-Learn.org, 2018, scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html.
- Singh, Shubham. "Social Media Users — How Many People Use Social Media in 2022." Demand Sage, 15 Apr. 2025, www.demandsage.com/social-media-users/. Accessed 27 Sept. 2025.
- Srivastava, Nitish, et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." Journal of Machine Learning Research, vol. 15, no. 1, 2014, pp. 1929–1958, www.cs.toronto.edu/~rsalakhu/papers/srivastava14a.pdf.
- "Social Media Advertising and Marketing Worldwide." Statista, 2024, www.statista.com/topics/1538/social-media-advertising/. Accessed 22 Sept. 2025.

Ukraine's Economic Recovery: Sectoral Pathways to Post-War Growth By Alex Shymkiv

Abstract

Ukraine's economic survival during war is largely dependent on its ability to maintain stability and build the foundation for long-term growth. This paper examines how agriculture, defense technology, and natural resources together form the backbone of Ukraine's wartime and post-war recovery strategy. Drawing on theories of post-conflict reconstruction, economic diversification, and wartime investment behavior, it explores how these sectors embody the theoretical framework that translates resilience under conflict into lasting growth. Agriculture provides stability through high export revenue; defense technology fosters rapid innovation, emerging as a new pillar of the global economy; and natural resources hold immense long-term potential rooted in the country's vast and largely untapped reserves. While each sector faces distinct challenges, from regulatory and legal barriers to limited investment and infrastructure damage, their combined development represents Ukraine's most viable path toward sustainable recovery and enduring economic prosperity.

Keywords

Economics, Development Economics, Foreign Direct Investment, Post-Conflict Reconstruction Investment Climate

Introduction

Not all battles are won through weapons; even when it comes to a country's survival, that survival often depends on far more than just military strength. For Ukraine, the ability to keep its economy functioning is as critical to survival as its ability to defend its territory from invaders. The country's GDP suffered a severe drop in the first year of the invasion, plummeting by almost one-third of its pre-war value (European Parliament). While GDP has been slowly recovering since, Ukraine still faces major economic struggles as the inflation rate remains high at 14.1% as of July 2025 (NBU), foreign direct investment has collapsed by around 97%, and 13% of the country's housing infrastructure is destroyed or damaged according to the World Bank. Still, signs of resilience have begun to emerge as targeted recovery efforts have taken shape across multiple industries. Yet, no sector can hold this burden alone.

Strategic Sectors Overview

Economic recovery during wartime requires a balance between short-term resilience and future-oriented investments, as well as the ability to innovate and adapt to rapidly changing conditions, because as historian A. J. P. Taylor said, "War has always been the mother of invention." Short-term resilience comes in the form of stability, a foundational element highlighting the country's immediate ability to sustain production and employment under constant disruption. Without stable revenue streams, even the strongest technological progress, aid, and reform wouldn't be effective or sustainable. Innovation, in this case, is equally

important, as it allows the economy to adjust to wartime constraints, develop new industries, and maintain competitiveness in the long run. Finally, future-oriented investments are essential as they lay the groundwork for post-war modernization and integration into global markets. Together, these three principles form the guiding framework for how Ukrainian economic recovery should be structured. They clarify not only what sustains the country now, but what will propel it forward once the war ends. Three specific sectors of Ukraine's economy stand out in providing complementary strengths across all of these areas.

First is the agricultural economic sector, which serves as the heartbeat of the Ukrainian economy accounting for approximately 10% of its GDP and over 40% of its export revenues – a stable source of revenue for the country (Régnier and Catallo). Its utmost importance was underscored specifically, when its stability was undermined at the start of the Russia-Ukraine war, as Russia imposed a naval blockade on Ukrainian Black Sea ports and mined shipping lanes, resulting in a world shortage of grain and dramatic price spikes. With Ukraine facing a potential loss of nearly 6 billion USD in grain export revenue (Reuters), and the world in need of this grain, the international community pushed for The Black Sea Grain Initiative. The initiative reopened maritime routes, vital for countries in Africa that had relied heavily on Ukrainian wheat exports before the war (Council on Foreign Relations), and for Ukraine as well, highlighting the utmost importance of the agricultural sector to the stability of the Ukrainian. Moving on, defense technology is the most dynamic of these sectors, one that rapidly scaled as the war began, reflecting Ukraine's innovation under pressure. Throughout the last 3 years of the War, the Ukrainian military has strengthened significantly, and while most of it can be attributed from high amounts of aid for the U.S. and EU, its innovation in the UAV industry is most remarkable, as the industry turned from simple battlefield necessity into entrepreneurial and technological opportunity. The country has emerged as a defense tech incubator, allowing companies to not only prototype and test their military technology, but also actively see its application in real time combat. Many international investors have already begun to commit capital into Ukrainian drone firms and military technology in general – contributing to a defense sector growth of over 350% since 2022 (Kuzio). Finally, Ukraine's natural resources sector, which has been getting a lot of attention recently due to the U.S.-Ukraine minerals deal, is the sector with the highest long term potential, as Ukraine is estimated to include two-thirds of Europe's rare earth reserves including not just the pricey oil and natural gas, but also natural resources like lithium, titanium, cobalt, and graphite (ifo Institute). These are critical for batteries, clean energy technologies, and high-tech manufacturing – industries that are expected to shape the future of the world.

Despite their promise, each of these sectors also faces several constraints that limit their ability to independently sustain the economy: legal and scaling barriers in agriculture, buyer monopsony in defense technology, and long and limited development timelines in the natural resources sector. That is why only the combination of short term revenue and necessary stability of the agricultural sector, military innovation through the defense tech sector, and unlimited long term potential of Ukraine's largely untapped natural resources is capable of forming a balanced and resilient foundation for Ukraine's economic recovery. To better understand how these sectors

function under wartime conditions and specifically how they can help Ukraine's recovery, it is necessary to examine the key economic theories that analyze economic recovery in conflict-affected countries, as well as further explore the sectors themselves.

Literature Review and Theoretical Context

Ukraine's wartime economy should specifically be analyzed through three major analytical frameworks like post-conflict reconstruction, economic diversification, and investor behavior, which help us explain how different economic forces interact under conditions of conflict and instability. The first important framework is post-conflict reconstruction theory, which focuses on how nations rebuild institutions, infrastructure, and production capacity after destruction, emphasizing the role of both foreign aid and domestic reform. Next is the concept of economic diversification in fragile states, which highlights the importance of developing multiple strong industries rather than relying on a single one, reducing vulnerability to sudden shocks or instability. The final framework to look at, wartime investment behavior, examines ongoing conflicts effects on investor confidence and behavior. Together, these frameworks underscore that sustainable recovery is grounded in resilience and adaptability, requiring strong institutions, diversified economy, and most importantly, strong FDI despite uncertainty. This understanding thus should be applied to Ukraine's agricultural, defense tech, and natural resources sector to assess how each contributes to, and is limited by, these broader economic dynamics.

Post-War Reconstruction Models

Post-war reconstruction literature emphasizes that successful economic recovery depends on both external aid and internal systems like strong courts and financial institutions that can further encourage investment and innovation. According to Kocherov et al., Ukraine's economic recovery will rely on this dual framework, heavily dependent on the combination of international support from the EU and the U.S., with reforms in domestic laws and governance. Ultimately, lasting recovery will hinge on Ukraine's ability to convert this external assistance into long-term institutional strength not just to please the European Union, but to build genuine economic sovereignty capable of sustaining growth on its own. Similarly, Kulikov et al. highlight innovation-driven projects and, more importantly, partnerships between private and state institutions as the cornerstone of sustainable reconstruction, aligning market efficiency with national development goals. Expanding on Kocherov et al.'s argument, Kulikov et al. assert that these collaborations allow for a transfer of expertise, technology, and capital, fostering an ecosystem of co-development, and eliminating one-sided dependence. Pidorycheva extends both of these arguments by looking at previous post war recovery models, drawing parallels to the Western European post-WW2 economic recovery through the Marshall Plan. She argues that Ukraine must replicate parts of these historical frameworks, while adapting them to modern realities of the 21st century by leveraging private innovation as the core engine of growth. In her analysis, sustainable recovery depends on empowering domestic entrepreneurs and integrating

them into global markets, using targeted international aid as the main catalyst. Overall, these studies collectively illustrate that Ukraine's recovery depends on the government's ability to effectively channel international support into high-potential sectors while fostering innovation, deregulation, and private development within them to drive self-sustained growth amid ongoing conflict. These insights also clarify why agriculture, defense technology, and natural resources stand at the center of Ukraine's recovery, as together, they reflect the perfect intersection of external aid, institutional reform, and private innovation needed to rebuild a resilient economy.

Economic Diversification Under Conflict

Economic theory also points to diversification as a key factor in resilience during and after conflicts. Venables, in his article *Using Natural Resources for Development: Why Has It Proven So Difficult?*, warns that overreliance on a single sector – especially in resource-rich economies – can lead to volatility, institutional fragility, and the “resource curse.” The resource curse refers to the paradox in which countries with abundant natural resources often experience slower economic growth and weaker governance, as wealth from resource exports crowds out investment in other productive sectors. In Ukraine's case, this risk is magnified by war-related disruptions that can instantly halt production or export activity in concentrated industries as previously we saw with the agricultural sector. Such dependence must be avoided in the post-war natural resources sector if Ukraine wants to successfully rebuild its economy and ensure sustainable growth. Furthermore, empirical research from Sukadi shows that economies with more diversified export structures suffer smaller output declines and recover faster following shocks. In practical terms, this means that countries investing in multiple productive sectors build flexibility into their economies, and don't rely on a singular sector to provide the necessary short-term stability, innovation, and long term potential. Diversifying across agriculture, defense technology, and natural resources sectors mitigates this vulnerability by spreading investment and creating overlapping value chains.

Wartime Investment Environment

Ukraine is an extremely complex investment environment defined by both opportunity and constraint. As previously mentioned, on one hand, the country is experiencing a brutal war that is actively harming its economy and has displaced over 10.6 million Ukrainians (UN in Ukraine), while on the other hand, the war has allowed for new forms of sectoral innovation to be born. This environment has created a paradox where wartime necessity drives innovation, even as economic restrictions tighten around investors and firms. As Kostruba explains, this conflict has required the state to impose many regulatory measures and capital controls that limit foreign profit repatriation to keep the economy afloat, essentially leaving investors with no choice but to reinvest their profits in Ukraine, unable to extract them. This pattern is clearly visible in recent data, where reinvested profits accounted for roughly 75% of total Ukraine's FDI inflows (Kyiv Consulting), allowing the country to record about 4.4 billion USD in total FDI (2.5% of GDP) after an almost complete pause in 2022 (OECD). Essentially, Ukraine's

investment landscape is caught between necessity and uncertainty as investors continue to reinvest not because of confidence, but because withdrawal is impossible. Long-term recovery will depend on Ukraine's ability to convert this passive investment into sustainable growth, because, as Teres notes, the missile strikes, infrastructure vulnerability and specifically legal uncertainty continue to deter new large-scale greenfield investment. While it is impossible to fully guarantee investor security or protect assets while Ukraine is still actively at war, progress in the legal and regulatory sphere has slowly begun to take shape. For example, in June of 2025, the Ukrainian Parliament passed the new PPP law designed to modernize infrastructure investment rules, align them with EU standards, and simplify legal frameworks (CMS Law-Now). Although the effectiveness of these recent reforms remains to be seen, it is clear that the Ukrainian economy requires even more transparent and lenient regulations in order to truly attract sustainable long-term investment.

Sectoral Analysis

While theoretical models explain the principles behind recovery and investment, their real test lies in how they apply to Ukraine's major economic sectors. This section explores how agriculture, defense technology, and natural resources can each uniquely contribute to Ukraine's economic recovery, revealing both opportunities and persistent vulnerabilities within them.

Agriculture

As previously mentioned, agriculture remains one of Ukraine's most strategically important economic centers, providing fiscal stability and export strength. Yet, surprisingly, it continues to chronically receive only a marginal share of external investment despite immense resource potential, attracting only 6% of Ukraine's total FDI inflows as of 2024 (UkraineInvest). This underinvestment is largely due to Ukraine's land market institutions and property-rights regulations like the 2021 Land Reform Law. As Onegina and Vitkovskyi note, the reform opened farmland purchases only to Ukrainian citizens and legal entities, explicitly excluding foreign ownership. This rapidly forced international agrobusinesses interested in investing in the sector to rely solely on leasing arrangements or establish subsidiary companies registered under Ukrainian ownership.

According to data from the European Parliament, Ukraine's large agricultural operators lease between 3 and 4.3 million hectares, roughly 7-10.5% of the country's total 41 million hectares of arable land. These restrictions have significantly complicated investment procedures and collateral opportunities in the sector, directly discouraging long-term FDI. These legal and structural limitations are further compounded by inefficiencies in state regulations, as emphasized by Dziurakh et al. He even states that Ukraine's agricultural sector's inconsistent policy enforcement, excessive bureaucratic oversight, and fragmented governance, make the sector one of the most administratively challenging investment environments in Eastern Europe. This is reflected in the sector's, as previously stated, extremely low investment composition, relying solely on reinvested earnings rather than foreign entrants (Puzikova).

Beyond these institutional obstacles, the war itself has brought large-scale physical destruction and supply chain paralysis, compounding existing market inefficiencies. As Teixeira et al. note, the ongoing war has destroyed major farmland and logistics chains, and left vast regions mined, reducing national harvest volumes by over 30% as of 2023. Damaged rail networks, fluctuating Black Sea port access, shifting routes through the Danube, and increasingly common farmer blockades, like the consistent Polish border protests that halt grain exports for weeks (Reuters), have significantly increased transaction costs and uncertainty. The European Parliament identifies demining and corridor protection as explicit prerequisites for any large-scale agricultural investment, noting that without secure infrastructure, even domestic investors face limited returns on capital. Thus, Ukraine's world-class factor endowments like fertile black soil, established production expertise, and comparative advantage in grain and oilseeds, remain undermined by the absence of stable market access and predictable security conditions.

Without renewed investment in agriculture, Ukraine risks losing its single most dependable source of foreign exchange and fiscal stability – the very foundation upon which its sustainable post-war economic recovery depends. Ultimately, consistent with Kocherov et al. and Kulikov et al., Ukraine's agricultural recovery ultimately depends on the synergy between external capital and domestic reform. Dibrova et al. emphasizes the need for new state initiatives designed for the post-war period, including targeted subsidies, low-interest credit facilities, and crop insurance mechanisms that help lower production risks for both domestic and international participants. Foreign assistance and EU-aligned frameworks can supply liquidity and technology, but without transparent governance and enforceable property rights, those inflows will remain limited to reinvested profits rather than new entrants. To attract sustainable FDI, Ukraine must therefore change its land-ownership laws, while simultaneously completing demining and logistics reconstruction. In doing so, agriculture will be able to not only regain its role as a high and stable source of income, but also evolve into a sector capable of attracting substantial and stable foreign direct investment.

Defense Technology

The defense technology sector has emerged as Ukraine's fastest-growing and most dynamic industry since the start of the full-scale invasion, completely reshaping the country's wartime economy and investor landscape. As modern warfare has increasingly shifted toward drone and electronic combat, Ukraine has become both a producer and a live testing ground for next-generation defense innovation. The front lines have effectively turned into laboratories for rapid prototyping and testing of advanced electronic warfare tools in real combat conditions.

No Western companies can replicate this, which is why many are racing to secure partnerships and investment opportunities with Ukrainian producers. For instance, BAE Systems opened a local office in Kyiv in 2023 to co-produce artillery systems and explore drone manufacturing contracts, while Rheinmetall and Baykar have announced plans to invest over 200 million USD combined into joint production facilities (Lapaiev). According to The New Voice of

Ukraine, foreign participation in Ukraine's defense tech sector is upwards of 100 million USD, signaling growing global recognition of Ukraine as both a testing hub and a critical player in the emerging miltech sector.

As a result of this rapid collaboration between state initiatives and private innovation, production value in Ukraine's defense industry has surged twentyfold compared to pre-war levels, reaching 30.8 billion USD as of 2023 (Kuzmuk & Scarazzato). This transformation reflects what Kocherov et al. describe as "war-driven industrial mobilization," where necessity forces innovation to move faster and reach necessary production quicker. The creation of the Brave1 innovation cluster in April 2023 under Ukraine's Ministries of Digital Transformation, Defense, and Economy further illustrates this transformation. Having already disbursed approximately 52.7 million USD in grants to Ukrainian defense tech startups since its inception (Ukrainska Pravda), Brave1 connects developers directly with government procurement agencies and military units, helping startups secure faster access to funding while simultaneously accelerating development and innovation. The platform has already attracted hundreds of startup submissions and growing attention from international investors, many of whom see Brave as a direct gateway into Ukraine's rapidly expanding defense tech market.

The Test in Ukraine initiative is another key mechanism of cooperation, enabling foreign defense firms to test weapons systems in live conditions alongside Ukrainian troops (Reuters). This allows for the collection of real-time performance data, which demonstrates the effectiveness of these new technologies, strengthening investor confidence. The sector's growing appeal is further underscored by the 59 million USD raised by Ukrainian defense startups in 2024 (DealBook of Ukraine), the first significant wave of foreign venture capital entering Ukrainian military technology. However, this rapid scaling has also exposed deep structural constraints. As Kostruba notes, Ukraine's miltech sector still functions as a monopsony, where the government remains the dominant, and pretty much sole, buyer. This reliance on the Ukrainian government as a single buyer, however, exposes an even deeper problem. In 2023, total defense orders were valued at 10.5 billion USD (Defence Industry Europe), covering less than half of the sector's total production capacity, forcing many firms to operate below 50% production efficiency levels, with many lacking formal government contracts altogether according to the UCDI report published a year later.

This imbalance narrows profit margins, both distorting market stability and even discouraging long-term foreign investment. Irtysheva et al. argue that this buyer concentration undermines healthy competition, a challenge further compounded by wartime secrecy and limited foreign licensing. Several manufacturing sites in Kyiv and Dnipro have already been repeatedly targeted by Russian missile strikes that, if not completely disrupting production cycles, have certainly left potential investors concerned about operational security. In response, however, several firms have begun forming partnerships with NATO countries, most notably through co-production agreements with Poland, the Czech Republic, and the Baltic states, in order to gain access to secure production facilities and more predictable procurement environments (Litra).

Despite these structural limits, Ukraine's defense sector remains an unmatched engine of technological innovation that can scale up to meet both domestic defense needs and future export demand. It stands as both a product of necessity and a blueprint for the country's future recovery, paralleling Kulikov et al.'s point about war-time urgency accelerating modernization rather than paralyzing it. The challenge now, though, lies in institutionalizing this momentum by transitioning from wartime production to a sustainable, export-driven industry that can contribute to long-term GDP growth. Doing so would, even more importantly, also integrate Ukraine's defense tech ecosystem into NATO and EU supply chains, securing strategic and economic partnerships that can further contribute to Ukraine's post-war resurgence.

Natural Resources

Ukraine's natural resource sector serves as the missing piece to the country's post-war economic recovery puzzle, representing one of the country's most strategically valuable and forward-looking sources of growth. Ukraine contains over 20,000 identified deposits of 116 different minerals, including significant reserves of iron ore, graphite, coal, titanium, and lithium, as well as oil and natural gas (World Economic Forum). Ukraine's "raw materials heart of Europe" name fits it well, as the country holds roughly 20% of global titanium, 6% of Europe's lithium reserves, and about 5% of the world's total mineral resources in general (Liepins).

Yet, despite its vast reserves, the sector has yet to reach its full productive and investment potential, as years of prolonged conflict risks have constrained its development. Before the war, the mining sector accounted for roughly 10% of the country's GDP and nearly 33% of its total export revenues (World Economic Forum). The Donetsk and Dnipropetrovsk regions, once vibrant industrial centers, together produced over 80% of Ukraine's steel and iron ore output, while also hosting some of the world's largest titanium and coal reserves (Conflict and Environment Observatory).

However, much of this production base has been severely disrupted by armed conflicts in eastern Ukraine since 2014, which saw Russia occupy large portions of Ukraine's industrial and mining areas, and was even further exacerbated by the 2022 full-scale invasion. As of 2024, the Ministry of Ecology and Natural Resources reported that approximately 40% of Ukraine's mineral deposits lie in temporarily occupied territories, including the lithium-rich Dobra and Shevchenkivske fields in Zaporizhzhia and Donetsk oblasts (Duggal & Ali). This geographic reality introduces high levels of uncertainty to any long-term investment projections, as the total value of Ukraine's natural resources, and the FDI they might attract, will ultimately depend on how much of its eastern and southern territories the country retains.

Despite these constraints, Ukraine's resource endowment has not gone unnoticed by its Western partners. In December of 2023, Kyiv and Washington signed the U.S.-Ukraine Critical Minerals Partnership deal, designed to co-develop rare-earth projects in Ukraine while channeling the resulting revenues into reconstruction funds (Reuters). Simultaneously, the European Commission expanded Ukraine's participation in the EU Raw Materials Alliance under the Critical Raw Materials Act, integrating Ukrainian mining into Europe's industrial

ecosystem (ERMA). Consequently, analysts estimate that full integration could attract upwards of 15 billion USD in FDI by 2033 if security conditions stabilize, and necessary institutional reforms take place (MarketVector).

Still, turning potential into increased production and revenue is extremely challenging due to the wartime environment's amplified traditional investment risks. Same as for agriculture, demining needs, infrastructure damage, and security threats make exploration both expensive and unpredictable. The destruction of key steel plants in Mariupol and mined maritime routes have severed traditional export corridors forcing producers to depend on congested western rail lines and Danube ports, sharply increasing logistics costs. Repeated Russian strikes on mining and processing sites have further discouraged insurers and private lenders, leaving most foreign involvement, like those from the U.S. and EU, limited to simple framework agreements and feasibility studies rather than actual capital investment. Even before Russia's 2022 full-scale invasion, major energy and mining investors had struggled to maintain operations under unstable conditions. A defining example is Shell's withdrawal from the Yuzivska shale gas project in 2015, a multibillion-dollar venture that was expected to help Ukraine reach energy independence (Rachkevych). Shell's exit, prompted by escalating conflict in Donbas, illustrated that large-scale resource projects require not only capital but also time, territorial control, and predictability, in order for investors to commit long-term, all of which remain fragile under ongoing conflict. Development timelines for lithium or titanium mines can exceed seven to ten years from discovery to full-scale extraction, meaning any serious foreign commitment will hinge on long-term peace guarantees and, once again, reform.

Recognizing these structural barriers, Ukraine has already undertaken several institutional reforms aimed at restoring investor confidence. The Prozzoro digital auction platform now manages subsoil license tenders, introducing transparency into what was once a shady process often criticized for corruption and oligarch favoritism (World Justice Project). The revised 2023 Subsoil Code has simplified exploration permits, closely aligning with environmental and licensing standards of the EU (Yermolenko). Additionally, the government has begun mapping critical minerals digitally and offering public data through the Unified State Geological Fund, which has been praised by the EBRD and IMF as a key step toward improving risk assessments for investors (Ukrainian Geological Survey). While these reforms are improving investor perceptions, actual implementation remains uneven as fewer than a quarter of exploration licenses have advanced to production largely due to the previously mentioned issues (Dychkovskyi et al.).

Despite these challenges, the sector's long-term promise is undeniable. As Ukraine's reconstruction gains momentum, the natural resource industry could evolve from a vulnerability into a strategic asset. An asset capable of not only driving financial recovery, but also supporting Europe's shift away from Russian and Chinese supply chains, strengthening Ukraine's geopolitical importance (Dentons). Analysts further project that, with consistent governance and foreign technical assistance, the sector could generate annual export revenues in the mid-tens of billions of USD by the 2030s, far surpassing pre-war agricultural export levels (Reuters).

Ultimately, Ukraine's natural resources sector is filled with both immense challenges and extraordinary potential for FDI and revenue growth. With stable conditions and credible investor protection guarantees, Kyiv could transform the sector from dormant wealth into a driving force of reconstruction and a foundation for Ukraine's long-term economic prosperity.

Conclusion

Ukraine now stands at a pivotal moment where recovery demands more than just reconstruction, but a structural transformation capable of sustaining prosperity beyond the war. The analysis of the agriculture, defense technology, and natural resources sectors shows each area's unique contributions to this process. Investment in agriculture will provide short-term stability, the rapidly emerging miltech sector, driven by innovation, will capitalize on wartime innovation, and FDI in natural resources will anchor the country's long-term resurgence.

Agriculture remains Ukraine's stabilizing force, yet its potential is constrained by weak property laws and disrupted logistics. Without reform and infrastructure recovery, it risks remaining a simple lifeline rather than a strong catalyst for growth. Defense technology, meanwhile, stands as Ukraine's most dynamic industry, exemplifying war-driven industrial mobilization. It proves that necessity can foster innovation, but sustained success will depend on shifting from state dependence to export integration with NATO partners. The natural resources sector holds the greatest long-term promise. With vast reserves of rare earth metals, it could anchor Ukraine's economic modernization and integrate the country into Western supply chains. However, realizing this potential will require territorial security as well as decades of sustained investment and transparent regulation.

Together, these findings reaffirm that Ukraine's post-conflict recovery depends on diversification across industries. Its challenge now is to not only rebuild what was lost, but to institutionalize transparent reform and innovation, building a resilient economy capable of attracting lasting FDI.

Works Cited

- European Parliament. “Two Years of War: The State of the Ukrainian Economy in 10 Charts.” 2024,
[https://www.europarl.europa.eu/RegData/etudes/BRIE/2024/747858/IPOL_BRI\(2024\)747858_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/BRIE/2024/747858/IPOL_BRI(2024)747858_EN.pdf).
- National Bank of Ukraine. “NBU July 2025 Inflation Update” 11 Aug. 2025,
<https://bank.gov.ua/en/news/all/komentar-natsionalnogo-banku-schodo-rivnya-inflyatsiyi-v-lipni-2025-roku>.
- World Bank. “Foreign Direct Investment, Net Inflows (BoP, Current US\$) – Ukraine.” 2024,
<https://data.worldbank.org/indicator/BX.KLT.DINV.CD.WD?locations=UA>.
- World Bank. “Updated Ukraine Recovery and Reconstruction Needs Assessment Released.” 25 Feb. 2025,
<https://www.worldbank.org/en/news/press-release/2025/02/25/updated-ukraine-recovery-and-reconstruction-needs-assessment-released>.
- Taylor, Alan J. P. “The Struggle for Mastery in Europe, 1848–1918.” Oxford University Press, 1954, <https://archive.org/details/dli.bengal.10689.13016>.
- Régnier, Elsa and Catallo, Aurélie. “Ukraine’s Agriculture and the EU Accession Perspective.” Institute for Sustainable Development and International Relations (IDDRI), 2024,
<https://www.iddri.org/sites/default/files/PDF/Publications/Catalogue%20Iddri/Etude/202406-ST0324-ukraine%20EU.pdf>.
- Reuters. “Ukraine Could Lose \$6 Billion in Grain Exports with Ports Blocked.” 21 Mar. 2022,
<https://www.reuters.com/business/ukraine-could-lose-6-bln-grain-exports-with-ports-blocked-2022-03-21/>.
- Council on Foreign Relations. “How Ukraine Overcame Russia’s Grain Blockade.” 26 Feb. 2024, <https://www.cfr.org/article/how-ukraine-overcame-russias-grain-blockade>.
- Kuzio, Taras. “Russia’s War Transforms Ukraine into a World-Leading Military Producer.” The Jamestown Foundation, 2025,
<https://jamestown.org/program/russias-war-transforms-ukraine-into-a-world-leading-military-producer/>.
- ifo Institute. “Ukraine Could Play Key Role in EU Supply of Raw Materials.” 23 Apr. 2025,
<https://www.ifo.de/en/press-release/2025-04-23/ukraine-could-play-key-role-in-eu-supply-of-raw-materials>.
- Kocherov, Mykhaylo; Dzhyhora, Olha; Dykha, Mariia; Lukianova, Valentyna; Polozova, Valentyna. “Mechanisms of Post-War Economic Recovery in Ukraine: The Role of the International Community.” *Economic Affairs*, vol. 68, no. 2, 2023, pp. 1311–1321,
<https://doi.org/10.46852/0424-2513.2.2023.35>.
- Kulikov, Petro; Aziukovskyi, Oleksandr; Vahonova, Oleksandra; Bondar, Olena; Akimova, Lyudmila; Akimov, Oleksandr. “Post-War Economy of Ukraine: Innovation and Investment Development Project.” *Economic Affairs*, vol. 67, no. 5, 2022, pp. 943–959,
<https://doi.org/10.46852/0424-2513.5.2022.30>.
- Pidorycheva, Iryna. “Post-War Recovery of Europe: Experience and Lessons for Ukraine.” *Journal of European Economy*, vol. 21, no. 2, 2022, pp. 170–187,
<https://doi.org/10.35774/jee2022.02.170>.

- Venables, Anthony J. "Using Natural Resources for Development: Why Has It Proven So Difficult?" *Journal of Economic Perspectives*, vol. 30, no. 1, 2016, pp. 161–184, <https://doi.org/10.1257/jep.30.1.161>.
- Sukadi, Emmanuel A. "Does Economic Diversification Foster Resilience to Crises? Empirical Investigation." SSRN, 2022, <https://ssrn.com/abstract=4273418>.
- United Nations in Ukraine. "UNHCR: After Three Years of War, Ukrainians Need Peace and Aid." 18 Feb. 2025, <https://ukraine.un.org/en/289506-unhcr-after-three-years-war-ukrainians-need-peace-and-aid>.
- Kostruba, Anatoliy. "Managing Foreign Business Operations in Ukraine in the Context of War." *Business Horizons*, vol. 68, no. 1, 2025, pp. 67–81, <https://doi.org/10.1016/j.bushor.2024.01.003>.
- Kyiv Consulting. "Investing in Ukraine 2025: What Investors Need to Know Before Investing in Ukraine in 2025." 2025, <https://kyivconsulting.com/insights/what-investors-need-to-know-before-investing-in-ukraine-in-2025>.
- Organisation for Economic Co-operation and Development (OECD). "OECD Economic Surveys: Ukraine 2025." 2025, https://www.oecd.org/en/publications/oecd-economic-surveys-ukraine-2025_940cee85-en/full-report/fostering-macroeconomic-stability-and-a-sustainable-recovery_24ed81a1.html.
- Teres, Yuliia. "Attracting Foreign Investments for the Development of the Ukrainian Economy: Wartime Challenges and Postwar Prospects." *Economy and Society*, 2023, <https://doi.org/10.32782/2524-0072/2023-51-51>.
- CCMS Law-Now. "Ukraine Overhauls Its PPP Regime, Facilitating Reconstruction and Investment." 7 July 2025, <https://cms-lawnow.com/en/ealerts/2025/07/ukraine-overhauls-its-ppp-regime-facilitating-reconstruction-and-investment>.
- UkraineInvest. "Agrifood Industry Overview." 2024, <https://ukraineinvest.gov.ua/en/industries/agrifood/>.
- Onegina, Viktoriya, and Yurii Vitkovskiy. "Investments and Land Reform in Agriculture of Ukraine." *Agricultural and Resource Economics: International Scientific E-Journal*, vol. 6, no. 4, 2020, pp. 187–210, <https://doaj.org/article/af609b2a32ce4493b1979f049ea91d02>.
- European Parliament. "Ukraine's agriculture and the EU accession perspective." 2024, [https://www.europarl.europa.eu/RegData/etudes/BRIE/2024/760432/EPRS_BRI\(2024\)760432_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/BRIE/2024/760432/EPRS_BRI(2024)760432_EN.pdf).
- Dziurakh, Yurii; Halanets, Volodymyr; Kulakovska, Tetiana; Pugachov, Mykola; Pugachov, Volodymyr. "Economic Assessment and State Regulation of the Agri-Industrial Sector in Ukraine: Solutions for Ensuring Food Security (Modeling and Analysis)." *Economic Affairs*, vol. 67, no. 4, 2022, pp. 777–786, <https://www.proquest.com/scholarly-journals/economic-assessment-state-regulation-agri/docview/2755159290/se-2>.
- Puzikova, Valentyna. "Foreign Direct Investment in Ukraine (Hannover Economic Papers No. 706)." *Leibniz Universität Hannover, Wirtschaftswissenschaftliche Fakultät*, 2023, <https://hdl.handle.net/10419/283157>.

- Teixeira da Silva, Jaime A.; Koblianska, Inna; Kucher, Anatolii. "Agricultural Production in Ukraine: An Insight into the Impact of the Russo-Ukrainian War on Local, Regional, and Global Food Security." *Journal of Agricultural Sciences (Belgrade)*, vol. 68, no. 2, 2023, <https://doiserbia.nb.rs/img/doi/1450-8109/2023/1450-81092302121T.pdf>.
- Reuters. "Polish Farmers Anger Ukraine with Border Blockade, Grain Spillage." 20 Feb. 2024, <https://www.reuters.com/world/europe/polish-farmers-step-up-protests-with-total-blockade-ukrainian-border-2024-02-20/>.
- Dibrova, Anatolii; Ilchuk, Mykola, Konoval, Ivan; Androsovykh, Ivan; Zanizdra, Artur. "State Support for Agriculture in Ukraine in the Post-War Period." *Agricultural and Resource Economics: International Scientific E-Journal*, vol. 9, no. 3, 2023, pp. 49–76, <https://are-journal.com/are/article/view/707>.
- BAE Systems. "BAE Systems Establishes Local Presence and Signs Agreements to Support Ukraine." 31 Aug. 2023, <https://www.baesystems.com/en/article/bae-systems-establishes-local-presence-and-signs-agreements-to-support-ukraine>.
- Lapaiev, Yuri. "Ukraine Ramps Up Defense Production with Increased International Cooperation." Jamestown Foundation, 17 Sept. 2025, <https://jamestown.org/program/ukraine-ramps-up-defense-production-with-increased-international-cooperation/>.
- The New Voice of Ukraine. "Investors Bet on Ukraine's Combat-Tested Tech." 13 Aug. 2025, <https://english.nv.ua/nation/who-is-investing-in-ukrainian-defense-tech-and-why-analysis-by-nv-business-50536830.html>.
- Kuzmuk, Kateryna; Scarazzato, Lorenzo. "The Transformation of Ukraine's Arms Industry amid War." SIPRI, 2025, <https://www.sipri.org/commentary/topical-backgrounder/2025/transformation-ukraines-arms-industry-amid-war-russia>.
- Ukrainska Pravda. "Brave1 Defence Tech Cluster to Issue Grants for Production of Explosives." 24 Jul. 2025, <https://www.pravda.com.ua/eng/news/2025/07/24/7523191/>.
- Reuters. "Ukraine Offers Its Front Line as Test Bed for Foreign Weapons." 17 Jul. 2025, <https://www.reuters.com/business/aerospace-defense/ukraine-offers-its-front-line-test-bed-foreign-weapons-2025-07-17/>.
- The Defender Media. "DealBook of Ukraine 2024: Investments in Ukrainian Defence Tech Surpass \$59M." 3 Mar. 2025, <https://thedefender.media/en/2025/03/dealbook-of-ukraine-2024/>.
- Defence Industry Europe. "Ukraine's defence industry is set to reach USD 35 billion production capacity in 2024." 2 Mar. 2025, <https://defence-industry.eu/ukraines-defence-industry-set-to-reach-usd-35-billion-production-capacity-in-2024/>.
- Ukrainian Institute for the Future & Ukrainian Council of Defence Industry. "Ukrainian Institute for the Future and the Ukrainian Council of Defence Industry presented a report on the

results of the defence industry in 2024.” 17 Mar. 2025,
<https://ucdi.org.ua/en/news-en/report-on-the-results-of-the-defence-industry-in-2024/>.

Irtysheva, Inna; Kramarenko, Iryna; Sirenko, Ihor. “The economy of war and postwar economic development: World and Ukrainian realities.” *Baltic Journal of Economic Studies*, vol. 8, no. 2, 2022, <https://ideas.repec.org/a/bal/journal/2256-074220228210.html>.

Litra, Leo. “Defence co-production of Ukraine: from wartime necessity to strategic industry.” New Europe Center, Sept. 2025,
https://neweurope.org.ua/wp-content/uploads/2025/09/Defence-co-production_eng_web.pdf.

World Economic Forum. “The future of critical raw materials: How Ukraine plays a strategic role in global supply chains.” 9 Jul. 2024,
<https://www.weforum.org/stories/2024/07/the-future-of-critical-raw-materials-how-ukraine-plays-a-strategic-role-in-global-supply-chains/>.

Liepins, Aigars. “Ukraine’s resources: Critical raw materials.” NATO Energy Security Centre of Excellence (ENSEC COE), 10 Dec. 2024, <https://www.enseccoe.org/publications/ukraines-resources/>.

Conflict and Environment Observatory. “The Mineral Wars: How Ukraine’s Critical Minerals Will Shape the Battlefield of the Future.” *Horizons: Journal of International Relations and Sustainable Development*, 2025, issue no. 29.
<https://www.cirsd.org/en/horizons/horizons-winter-2025-issue-no-29/the-mineral-wars>.

Duggal, Hanna; Ali, Mariam. “Mapping Ukraine’s Rare Earth and Critical Minerals.” *Al Jazeera*, 28 Feb. 2025,
<https://www.aljazeera.com/news/2025/2/28/mapping-ukraines-rare-earth-and-critical-minerals>.

Reuters. “Ukraine Revamps Minerals Sector, Eyes Billions in Investment under U.S. Deal.” Reuters, 27 May 2025,
<https://www.reuters.com/world/europe/ukraine-revamps-minerals-sector-eyes-billions-investment-us-deal-2025-05-27/>.

European Raw Materials Alliance (ERMA). “The Start of Long-Standing Cooperation: EU and Ukraine Sign a Strategic Partnership on Raw Materials and Welcome a New Member of ERMA.” 13 Jul. 2021,
<https://erma.eu/the-start-of-long-standing-cooperation-eu-and-ukraine-sign-a-strategic-partnership-on-raw-materials-and-welcome-a-new-member-of-erma/>.

Rachkevych, Mark. “Shell Pulls Out of East Ukraine Gas Exploration Project.” *Kyiv Post*, 11 Jun. 2015,
<https://archive.kyivpost.com/article/content/business/shell-pulls-out-of-east-ukraine-gas-exploration-project-390803.html>

World Justice Project. “Ensuring Transparency through Prozorro.Sale: The World’s Most Transparent Auction System.” 2022,
<https://worldjusticeproject.org/world-justice-challenge-2022/se-prozorrosale>

Yermolenko, Halina. “Changes to the Rules for the Use of Subsoil Have Entered into Force in Ukraine.” *GMK Center*, 5 Apr. 2023,
<https://gmk.center/en/news/changes-to-the-rules-for-the-use-of-subsoil-have-entered-into-force-in-ukraine/>

- Ministry of Environment and Natural Resources of Ukraine. “Ukrainian Geological Survey a modern service-oriented institution” 2023,
<https://www.geo.gov.ua/wp-content/uploads/presentations/en/ugs-modern-service-oriented-institution.pdf>
- Dychkovskiy, Roman; Saik, Pavlo; Sala, Dariusz; Cáceres Cabana, Edgar. “The Current State of the Non-Ore Mineral Deposits Mining in the Concept of the Ukraine Reconstruction in the Post-War Period.” *Mineral Economics*, vol. 37, no. 2, 2024, pp. 589–599,
<https://doi.org/10.1007/s13563-024-00436-z>
- Dentons. “Ukraine’s Critical Minerals: A Strategic Asset for Global Supply Chains.” 20 Aug. 2024, <https://www.dentons.com/en/insights/articles/2024/august/20/ukraine-critical-minerals>

To What Extent Can AI-Driven Predictive Tools Improve Agricultural Outcomes for Farmers in Indian States Vulnerable to Unpredictable Conditions? By Nikunj Garg

Abstract

Indian farmers have to deal with issues like delays in rain, an unforeseen insect outbreak, or a sharp decline in crop prices constantly. In addition to reducing crop yield, these circumstances cause many farmers to incur greater debt. The paper will investigate if artificial intelligence (AI) can increase the resilience of the Indian agriculture market to these threats. It will examine how AI can be used to forecast weather, optimize irrigation, and even identify crop illnesses early. Initial findings are encouraging with pilot projects indicating when to be alert about possible issues, providing advice on how to have more consistent harvests, and advising about more intelligent use of fertilizer and water. These improvements are not, however, as many farmers still don't have access to digital resources, reasonably priced internet, or the necessary training / literacy to use these tools. The study contends that although AI has the potential to significantly improve agricultural results, its effectiveness will rely on how governments can reduce divides and make innovative technology like AI available and usable to those who need it most.

Introduction

Agriculture continues to be the backbone of rural India, employing nearly half of the country's workforce and shaping the livelihood of millions. Despite this, farmers in India are routinely at the mercy of delayed monsoons, sudden droughts, unseasonal floods, pest infestations, or sharp swings in crop prices. These unpredictable conditions have made Indian agriculture one of the most vulnerable sectors in the world, with smallholders—who make up the majority—often bearing the heaviest losses.

In recent years, artificial intelligence (AI) has begun to emerge as a possible answer to these challenges. By looking at satellite data, weather data, and crop data, AI tools can develop to forecast rainfall patterns, optimize irrigation schedules, and predict crop yields. Projects such as Microsoft's AI Sowing App in Andhra Pradesh and other pilot programs suggest that technology like this can reduce uncertainty and help farmers make better decisions based on real data.

This paper looks at the extent to which AI-driven prediction tools can improve agricultural outcomes in India for farmers. It examines not only how effective these tools are in addressing unpredictable conditions but also what barriers, technical, economic, and social, and in the way of their broader adoption by and for farmers.

Literature Review

Indian agriculture has historically been susceptible to a variety of unpredictabilities, including erratic weather, pests and diseases, and climate. Weather patterns have become more unpredictable due to climate change, with heatwaves, floods, and drought occurring more

frequently and upsetting cropping cycles. Normal forecasting techniques consistently fail in offering farmers accurate, timely, or localized predictions. In order to provide more precise and location-specific predictions, this gap has led research into artificial intelligence (AI) models that can process large amounts of satellite data, crop data, and historical weather records. Based on initial research and trials, these technologies can and will decision-making about irrigation and planting schedules if implemented properly (Frontiers, 2025).

Farmers additionally have to contend with the issue of uncertainty brought on by social and economic upheavals. Farmers are unable to sell their produce because of events like the COVID-19 epidemic, transportation strikes, or random regulatory limitations on mobility and trade that drastically decreased demand in local markets and areas. Although these shocks are not the same as natural disasters, they both weaken planning ability and income stability and are circumstances in which the use of AI may be appropriate. Using consumption patterns, mobility data, and regional indicators, recent studies have investigated to see if predictive analysis might help in anticipating changes in demand. According to this research, even though AI development is still in its early stages, it could help farmers by warning of possible disruptions before they become more serious (WEF, 2025).

Findings / Analysis

Evidence from initial programs and early studies indicate that AI tools are already improving the way farmers are able to respond to unpredictable conditions. In Andhra Pradesh, for example, Microsoft and the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) created an AI-powered “Sowing App.” The tool analyzed available historical climate data about the region, soil conditions in the area, and satellite imagery to provide advise for farmers on the optimal sowing dates for their crops. According to reports, farmers using the developed app saw yield increases of 10 to 30 percent as compared to those relying on traditional schedules (Microsoft & ICRISAT, 2024). This suggests that even simple mobile-based AI tools that wouldn't require many resources beyond data inputs could make a tangible difference in managing the uncertainty caused by erratic weather in these regions.

Another area where AI has shown clear benefit is in water management. AI-driven irrigation models combine weather forecast data with soil moisture data to create precise schedules for irrigation that minimize water waste, reducing waste and improving the environmental impact. The World Economic Forum’s Saagu Baagu project reported that farms using these irrigation schedules developed by these systems in Telangana reduced water use by up to 15 percent while maintaining or improving crop yields (WEF, 2025). This is especially relevant as rising temperatures and uneven rainfall place additional stress on groundwater resources and when the increased water demands that would come with the use of AI on a wide scale are considered.

Beyond individual pilot projects, a broader analysis published in Frontiers regarding Sustainable Food Systems found that machine-learning weather models consistently outperform conventional meteorological predictions in short-term forecasting accuracy (Frontiers, 2025).

These gains enable farmers to better plan for heat waves, rainfall deficits, or unexpected weather events in general.

Overall, the findings show that AI does not eliminate unpredictability but provides farmers with a more informed basis for decision-making. The improvements seen in yield reliability, irrigation efficiency, and forecasting accuracy illustrate that AI can meaningfully strengthen resilience and improve the outcomes from farmers in the short term.

Discussion

However, while AI-driven tools offer new ways to reduce uncertainty in farming, their impact across India remains limited. Successful initiatives, such as Microsoft's Sowing App or the Saagu Baagu project, functioned in controlled settings with reliable infrastructure and expert oversight. What must be considered though is that in much of rural India, however, poor internet connectivity, high costs, and limited access to smartphones make such systems difficult to adopt.

Even in places where access exists consistently, farmers often face challenges interpreting AI recommendations due to literacy issues or conversational differences based on regional methods of communication. Many lack digital literacy or training, and the models themselves depend on data that are very frequently incomplete or outdated due to lacking administrative systems. Predictive systems trained on one region's data can also misrepresent conditions elsewhere, leading to inaccurate forecasts, especially when the difference in weather conditions in different areas of India are considered.

These constraints suggest that AI cannot, single handedly, transform Indian agriculture. Without investments in digital infrastructure, education, and administration systems allowing for consistent, localized data collection, its benefits will remain concentrated in small pilot regions rather than widely improving resilience to unpredictable conditions to the potential the technology has.

Conclusion

AI based predictive tools have shown that technology can really help farmers respond more effectively to India's agricultural conditions which feature unpredictable weather, regional difficulties, and drought. From weather forecasting to irrigation planning, basic predictive projects demonstrate real, measurable benefits in yield stability and resource use. Yet these successes remain limited in scale. Without the requisite tools like affordable internet access, proper training, and reliable local data, most farmers cannot use these tools effectively. AI can enhance decision-making but cannot replace the systemic changes agriculture in India requires. Its true value lies not in replacing human judgment, but in supporting it—making farming more informed, not fully automated, and progress more inclusive.

Works Cited

- Ajith, S. “Yield Prediction, Pest and Disease Diagnosis, and Soil Fertility Assessment Using AI.” *Journal of Applied Research in AI and Agriculture*, Springer, 2025.
<https://link.springer.com/article/10.1007/s44187-025-00338-1>.
- “Remote Sensing and Artificial Intelligence: Revolutionizing Agricultural Forecasting.” *Frontiers in Sustainable Food Systems*, 2025.
<https://www.frontiersin.org/journals/sustainable-food-systems/articles/10.3389/fsufs.2025.1551460/full>.
- Microsoft, and International Crops Research Institute for the Semi-Arid Tropics (ICRISAT). “AI Sowing App and Advisory Programs in Andhra Pradesh.” *Microsoft News India*, 2024.
<https://news.microsoft.com/en-in/features/ai-agriculture-icrisat-upl-india/>.
- University of Chicago. “AI Is Transforming Weather Forecasting and That Could Be a Game-Changer for Farmers around the World.” *UChicago Climate Research*, 2024.
<https://climate.uchicago.edu/insights/ai-is-transforming-weather-forecasting-and-that-could-be-a-game-changer-for-farmers-around-the-world/>.
- World Economic Forum. *Future Farming in India: A Playbook for Scaling Artificial Intelligence in Agriculture*. World Economic Forum, 2025.
https://reports.weforum.org/docs/WEF_Future_Farming_in_India_A_Playbook_for_Scaling_Artificial_Intelligence_in_Agriculture_2025.pdf.

Legal Loopholes in Theory and Practice: Historical Origins, Ethical Implications, and the Problems between Justice, Interpretation, and Exploitation By Elizabeth Wong

Introduction

The concept of legal loopholes reveals an important issue within modern legal systems: the divide between the letter of the law and its spirit. While loopholes may seem like minor technical oversights, their exploitation demonstrates how the law can be manipulated in ways that preserve formal legality while undermining the principles of justice.

The relevance of this issue is undeniable, especially in an age where corporate structures, financial markets, and even criminal prosecutions often hinge on the interpretation of highly complex regulations; loopholes function as both shields and weapons. This phenomenon raises critical questions about fairness, equality before the law, and the broader legitimacy of the justice system. This paper will discuss the historical context, theoretical ideologies, real-world examples, and empirical evidence provided by professionals.

1.1 Defining Legal Loopholes: Distinctions from Vagueness and Interpretation

In this section, this paper will clarify what constitutes a loophole and distinguish it from other similar terms such as legal vagueness and legal interpretation. Loopholes date many years back, from ancient attempts at dodging taxes to the structured codification of laws under Napoleon. Tracing their historical development reveals how the exploitation of legal gaps has led to reforms in legislation, leaving behind an impact that still influences modern law today.

As language cannot be conveyed with perfect clarity, all understanding of language is an interpretation. And as law is language, a differing interpretation of the law is inevitable. In this section, the definition of a legal loophole will be introduced, and how it differs from other similar terms such as legal vagueness and legal interpretation. This paper will establish a definition that will be referred to for the rest of the paper.

Legal loopholes do not have a concrete definition. Some view legal loopholes as gaps or ambiguities in the law that can be exploited to achieve a desired outcome that may not align with the original intent of the law (Nelson); and some others view it as a technicality that allows individuals or businesses to sidestep the scope of a law or restriction without directly violating it (Hayes). However, for this paper, it will be defined as gaps and ambiguous phrasing in laws which allows people to achieve goals that are regarded as unintended by the law. Some examples of this can include: tax avoidance schemes, contract loopholes, corporate liability dodging and more. With the complexity of the law, many people may confuse a legal loophole with other similar terms, and this paper will discuss some terms that it could be confused with: legal vagueness and legal interpretation.

Legal vagueness is generally discussed in philosophy of language in relation to open-textured terms with broad or indeterminate meaning (Christie 886). A loophole, by contrast, is precise in wording but leaves out a critical detail that can be exploited. Vagueness does not create such gaps; instead, it leaves space for multiple reasonable understandings. Legal

interpretation, meanwhile, refers to the process of assigning meaning to a legal text, where the norm is the meaning attributed by the interpreter (Yzaguirre & Morales 723). Unlike vagueness, which arises from the text itself, interpretation is the act of determining that meaning.

Defining these terms creates a clear framework for this paper, and building on this, the next section will examine the historical evolution of loopholes and codification, showing how their exploitation has shaped legislation and continues to influence modern law.

2.1 The Historical Evolution of Legal Loopholes and Codification

Historically, loopholes have played a significant role in shaping economies, influencing public policy, and driving legal reforms. These examples reveal systemic flaws that often lead to public criticism and societal change. Through understanding their historical significance and consequences, we can gain insight into the complexities of the fragile legal system, and understand how the law has progressed until today. This section will mainly focus on the history of loopholes, presenting how they developed to what they are now.

Today, one of the most significant forms of exploiting legal loopholes is tax evasion. However, even before concrete tax laws were established, people found ways to evade taxes thousands of years ago.

An example of one of the first cases of tax evasion was during the Egyptian period. Taxes were so heavy that many people fell into debt, especially during the Late Period of Ancient Egypt (c. 525-332 BCE). However, some poorer farmers found ways to evade their taxes by manipulating their weighted grain scale; people would sneak stones in the grain to meet the taxed weight for their fields. The problem grew so profuse, there were royal edicts issued telling people not to cheat the system. (Magazine & McMahon). Although this might not be a direct exploitation of loopholes, many people see this as one of the first ancient ways that pioneered the avoidance and bypassing of the law today.

This prevalent problem continued until 1804, when Napoleon introduced the Napoleonic Code — a comprehensive codification of laws intended to address and regulate all scenarios of human life. By establishing written law as the norm, the Code both unified and rationalised France's legal framework and set a global precedent for codified systems. However, with laws now fixed in written form, they became inherently more vulnerable to exploitation, as individuals could manipulate the precise wording to achieve outcomes never intended by the legislators.

After the French Revolution, powerful groups such as the manors and the guilds were destroyed, the secular power of the church was suppressed, and the provinces was transformed into subdivisions of the new national state. The lack of political unification demanded a new body of law that would be uniform for the entire state. And so, the Napoleonic Code was created, establishing on the idea of creating a purely rational legal system, free from historical prejudices and based on "sublimated common sense." Unlike previous laws rooted in tradition or monarchical authority, it was rooted in moral legitimacy, being grounded in its alignment with reason, marking the shift towards a rational approach to law (Encyclopedia Britannica).

This Napoleonic Code acted as a new legal framework for France, giving the post-revolutionary France its first coherent set of laws concerning four aspects: property, colonial affairs, family and individual rights (Onion et al.). Codification was not a new concept; however, the execution of codification was. This was never done before; execution quickly rose to popularity, causing it to be adopted by the whole of Europe, and branched out to the whole world, influencing the present legal system of written law.

With the establishment of the Civil Code, many of the distinctions between legal and illegal conduct were clarified; however, this also exposed the issues between moral principles and written rules. The next section will discuss the battle between two different understandings of how the law should be implemented in real cases.

3.1 Interpretation, Ideology, and the Rise of Loophole Exploitation

With the establishment of the written law, two schools of thought emerged; one side believed that the law should be empirical and practical, based on what can be directly perceived and measured, the other side proposed that law is intrinsically intertwined with abstract moral principles. This section will discuss the battle between formalists and realists, presenting court cases where both ideologies have been applied.

Both terms have accreted many meanings that each has become an all-purpose term, both of approbation and of disapprobation. Formalist can mean narrow, conservative, resistant to change, casuistic, callow, modest, reasoned; whereas "realist" can mean reductionist, manipulative, left-wing, progressive, humane, candid, mature, clear-eyed. These usages reflect the polemical character of so much writing about law.

On one hand, formalists are known to use deductive logic to derive the outcome of a case from premises accepted as authoritative. Formalism enables a commentator to pronounce the outcome of the case as being correct or incorrect, in approximately the same way that the solution to a mathematical problem can be pronounced correct or incorrect. On the other hand, realists decide the fate of a case so that its outcome best promotes public welfare in non-legalistic terms. A "realist" decision is more likely to be judged sound or unsound than correct or incorrect (Posner, 1986, pp.182).

However, many people have critiqued the formalistic approach to the law. They claim that the law is unable to cover the reality of the infinite number of scenarios. Human life is so complex that it surpasses written words, and some events may not be able to be addressed with the existing laws. When two people look at a paper, they never understand it the same way; when looking at the word "table", one might imagine a wooden table, while the other might imagine a plastic fold-out table. This hindrance in the understanding of one word could be catapulted when looking at statutes and legal texts. Lawyers and judges may use this to their advantage, bending the language to advance their own political agendas.

Some people contend that legal formalism is unable to cover the reality of many real-life cases. Human life is so complex that it surpasses written words, and some events may not be able to be addressed with the existing laws.

In 1889, (*Riggs v. Palmer*, N.Y. 1889), Francis B. Palmer made a will leaving most of his estate to his grandson, Elmer E. Palmer, while providing smaller legacies to his two other daughters. Worried that Francis would reconsider the will before he could inherit, Elmer poisoned Francis. Elmer was convicted of murder but still claimed the inheritance based on the will. Palmer's daughters filed a lawsuit to prevent Elmer from benefiting from the will, arguing that he should not profit from his crime. However, this complaint was dismissed and Elmer ended up receiving the inheritance anyway.

On the other hand, legal realism proposes the notion that all law derives from social interests, morality, and public policy. Legal realism is also thought of as a naturalistic approach to law; it should rely on empirical evidence and hypotheses that have been tested against the reality of the world, rather than rely on purely theoretical assumptions about the meaning of law (Cornell Law School).

However, some people argue that legal realism causes the law to be unpredictable instead of neutral, which puts the law's fairness, legitimacy and accountability into question. It could also be seen as a bet on a judge's personal biases and preferences.

In 2005, (*United States v. Booker* U.S.), the Supreme Court sentenced Freddie Booker, who was convicted by a jury of possessing at least 50 grams of crack cocaine. The Federal Sentencing Guidelines prescribed a sentence of 210 to 262 months based on the jury's findings. However, the judge found additional facts by a preponderance of the evidence, which increased the Guidelines range to 360 months to life, resulting in a 30 year sentence. The Seventh Circuit held that this violated the Sixth Amendment because the judge found facts that increased the sentence beyond the jury's determination. In a related case, *United States v. Fanfan*, the judge found additional facts that could have increased the sentence, but instead imposed a sentence based solely on the jury's findings. The Supreme Court was tasked with resolving whether the Federal Sentencing Guidelines, as applied, violated the Sixth Amendment and what remedy was appropriate. In the end, the Court affirmed the Seventh Circuit's decision in Booker's case and vacated and remanded Fanfan's case.

Though the projection of one's political belief could be dismissed, some individuals might take another approach to this, manipulating this fact to carry out illegal deeds. Legal interpretation has always been the scapegoat of offenders' illegal actions, and it has become quite common to see and catch people abusing this fact; And so, the discovery of loopholes made it easier for people to escape the consequences of their actions and cover up their wrongdoings. In the next section, this paper will discuss some famous cases of loopholes, how they have been executed, including the ethicality of the offender's actions and motivations and its subsequent influence.

4.1 Case Studies in Loophole Exploitation

To fully understand the depth of loopholes, this section will cover some famous cases of loopholes being carried out, the motivations behind the actions and the consequences. This

section will investigate the ethical issues related to these cases and will assess the underlying reasons for such actions.

Loopholes are problematic as they create an inequality between those who use it and those who do not, benefiting those with the resources to exploit them. These loopholes may also result in regulatory changes, potentially closing the loophole and affecting those who relied on it (Bamigbola).

4.2 The Revival of Dormant Legislation as a Tool for Policy Enforcement

A recent case of loopholes being abused is the current American President and the American Federal Government using ‘zombie laws’ — provisions that have not been enforced or invoked for decades but are somehow still in the books — to push his own political agenda about abortion towards the USA. The normal legal process to enforce this law would include Congress passing a new federal law to ban them. Instead, Trump and his team argue that the Comstock Act (1873) already bans mailing abortion pills. They claim that since the law has not been repealed, the government can enforce it now — even though it has not been used this way since the 1930s; with Jonathan F. Mitchell, a lawyer for Trump, saying “We don’t need a federal ban when we have Comstock on the books”. This circumvents Congress and allows Trump to shut down medication abortion without ever passing a new law.

The Comstock Act was originally intended to prevent the illegal provision of abortion pills in the 19th and 20th centuries, not to abolish them altogether. However, because this limitation was never clearly written into the law, Trump was able to exploit its ambiguity, using a legal loophole to reinvent the Act to support the passing of a new law, overriding an existing one.

Another loophole Trump wanted to use for mass deportations was the Alien Enemies Act. As stated above, Trump would need Congress to pass new immigration laws or expand deportation powers, which would face strong legal and political resistance from the public and Congress. So he dug up another law — The Alien Enemies Act (1798) — which allows the U.S. to deport or intern people from "enemy nations" during wartime. Trump could contend that the U.S. is facing a "foreign invasion" due to undocumented immigrants, manipulating and stretching the definition of “enemies” and “invasion” (Jong-Fast).

The law was clearly meant for wartime, yet Trump twists it through his legal interpretation of the terminology, using it for immigration enforcement instead, and with the definition of the “invasion of enemies” being broad enough to be able to apply to anyone, Trump uses this to enforce his own racial “right-winged” beliefs.

The issue with this is that the executive order sidesteps legal protections, allowing mass deportations — for many who are innocent — without giving people a chance to defend themselves in court, as the ACLU says “This measure would remove non-US citizens without any opportunity for judicial review.” Legal scholars have also argued that referring to illegal immigration as an ‘invasion’ may give Trump broad powers to deport individuals en masse or keep them in custody with no trial.

The policy relies on an authority that was last invoked in the Second World War to intern non-US citizens of Italian, German and Japanese descent — one of the most controversial episodes in American history; and by using a law with such a dark history, Trump's executive order raises serious ethical concerns about who might be targeted next under similar justifications. Trump's framing of anyone not fitting his idea of "radically American" suggests the possibility of wider application, potentially leading to racial profiling and unjust deportations of individuals with no criminal ties (Palma).

4.3 The Case of Mickey Barreto and Rent Stabilisation Law

Another person who exploited a legal loophole was Mickey Barreto, who managed to stay in a hotel for five years due to the New York Rent Stabilisation Law. This law was designed to protect tenants from excessive rent increases and unfair evictions while ensuring landlords receive regulated rent adjustments; However, this law also inadvertently allowed hotel guests to request leases for long-term residency in hotels. Under this law, hotel occupants can become "permanent tenants" as long as the hotel fits these criteria:

- In buildings of six or more units built between February 1, 1947 and December 31, 1973
- In buildings built before February 1, 1947
- In buildings with three or more apartments constructed or extensively renovated on or after January 1, 1974, with special tax benefits (Schneiderman, n.d.)

When Barreto and his partner, Hannan, discovered this law, they decided to try it out for themselves and used it to claim tenancy at the New Yorker Hotel.

After one night's stay, Barreto went to the front desk and handed the desk clerk a letter for the manager, requesting a six-month lease. The clerk called the manager, and the manager dismissed it. He hand-wrote a "three-page, handwritten affidavit that 'cited state laws, local codes, and a past court case.' arguing that the request for a lease made him a "permanent resident of the hotel." Then he added that "removal of his items amounted to an illegal eviction."

Barreto immediately went downtown to the New York City Housing Court, where he filed a suit against the hotel. However, judging by the unserious nature of the case, no one showed up for the hearing a few weeks later, leading the judge to order the hotel "to restore petitioner to possession of the subject premises forthwith by providing him with a key." Barreto, Hannan, and the judge's ruling in hand, were back in room 2565 a few days later. That's when they noticed that the judge's ruling made no mention of the hotel providing a lease, did not mention rent of any kind, and put no limit on the length of their stay, Leaving Barreto and Hannan to stay in the New Yorker Hotel until July 2023 without ever paying rent.

With his newfound success, Barreto pushed further. When he was granted a "final judgment of possession" in a court order, he interpreted it as granting him ownership over the hotel. He tried to register a deed for Room 2565 and then for the entire hotel building, eventually

succeeding in May 2019 when he was listed as the owner. Despite attempts by the hotel's owners to evict him, they could not provide evidence that the hotel was exempt from housing law.

Barreto, emboldened by his "ownership," demanded a share of the hotel's profits, tried to collect rent from tenants, and even asked for the hotel's accounts to be transferred to him. However, the hotel owners eventually managed to prove they had offered him leases, which he refused, and had tried to evict him. A judge ordered his eviction, but Barreto managed to re-register the deed under his name again, costing the hotel \$2.9 million in additional taxes due to a lost tax exemption (Hopkins Centrich PLLC).

In the end, Mr. Barreto was charged in February with 24 counts, including 14 felony fraud counts, in what the Manhattan district attorney's office said was a years-long criminal scheme to claim ownership of the New Yorker Hotel. However, Barreto was found unfit to stand trial due to mental illness and drug addiction. He faces criminal charges, but if he does not find inpatient psychiatric care, he may be hospitalized. (Haag)

4.4 Google's Double Irish Tax Avoidance Scheme

In the US, one of the most widely discussed cases of tax avoidance is Google's double Irish tax arrangement. Whilst Google reportedly earned over €22bn of sales revenue in 2016 from their Irish-registered company, Google Ireland Limited, they only paid €47m in tax (Bowers).

The Double Irish scheme involved a corporation setting up a subsidiary in Ireland but moving its headquarters to a tax haven like Bermuda, Cayman Islands or Seychelles, low-income countries where foreign companies were not taxed. The Irish holding company owned the IP rights, while a Dutch holding company licensed these rights and sold the product or service in high-tax countries. Profits or royalties then flowed back to the Irish company, which ultimately transferred them to the letterbox firm in the tax haven. Through this scheme, multinationals used these loopholes to shift profits to tax havens without paying tax. By registering "letterbox companies" in Ireland but moving their tax seat abroad, they created so-called "stateless income", criticised as immoral tax trickery harming ordinary taxpayers (FINEXITY Blog).

However, with mounting pressure from the OECD, EU, and U.S., in October 2014, Irish Finance Minister Michael Noonan announced the phaseout of the Double Irish as part of the government's 2015 budget. The government modified the Irish tax code such that all companies registered to do business in Ireland must be tax residents in the country. The law went into effect for new entrants on January 1, 2015, and the government gave existing companies based in Ireland and managed in another jurisdiction a five-year window to comply. This meant that Google could no longer utilise the arrangement starting in tax year 2020 (Organ & Samarakoon 2).

After reviewing real-life cases, interviews were conducted with legal professionals to discuss these loopholes. In the next section, the findings will be presented, showing a model of the varying opinions on loopholes.

5.1 Empirical Findings

This section presents findings from interviews conducted with legal professionals, including lawyers, judges, and legal academics on the subject of legal loopholes. The interviews explored their experiences, ethical views, and perspectives on the broader impacts and potential solutions related to loopholes in the legal system. Four key areas guided the inquiry: the general prevalence and origins of loopholes, their ethical implications, their effects and impacts on society and justice, and proposed reforms.

This exploratory study employed a qualitative interview methodology to investigate perceptions of legal loopholes among legal professionals and academics. A total of 18 participants, comprising practising lawyers, law professors, and law students. The primary objective of the questionnaire was to gather nuanced, experience-based insights rather than statistical generalisations.

Additionally, some participants responded orally during live interviews, while others provided written answers. Participants who were present for an oral interview were asked follow-up questions that were posed spontaneously in response to individual answers, meaning each interview varied slightly in content. However, the study's exploratory nature and the small, non-random sample mean that the findings are not representative of the wider legal profession, limiting the scope for generalisation. This interview serves as a preliminary foundation upon which further research by legal professionals can be built. The insights gathered here are intended to prompt deeper scholarly inquiry and informed debate, rather than to offer definitive conclusions.

5.2 General Questions

Participants offered differing interpretations of questions; some framed loopholes as necessary instruments within an adversarial system, while others saw them as distortions that undermine the law's intent. In this subsection, I have asked participants simple questions about the existence and prevalence of loopholes. Firstly, everyone agreed that legal loopholes are pervasive in the legal system; however, the reasons for their existence are torn. One lecturer said: *"[Legal loopholes] are pretty much common, it's everywhere, because firstly, it's natural language and natural language always has that ambiguity."*

This lecturer highlights that the very foundation of law — natural language — is inherently flawed. Because words could be vague or interpreted in multiple ways, loopholes become inevitable. However, she further states that the role judges played should have been to apply their interpretation of the situation and prohibit the abuse of these loopholes. In contrast, an international rights lawyer took a different approach and argued that loopholes are intrinsic to legal interpretation and adversarial practice, she said: *"Making arguments or making submissions ... are essentially based on legal loopholes ... All legal cases are predominantly disputes, and therefore there will be a side that tries to work with the law, and another side tries to find loopholes. So I think essentially, in the broader understanding, there are always loopholes in the law. No matter how well-drafted a law is, no matter how long the law has been in place,*

no matter how well-functioning a certain democracy or legal system is, I think the law will have numerous, numerous loopholes."

Here, she suggests that loopholes are often deliberate tools used in legal arguments. It paints the system as one where clever manipulation can outweigh intent. She suggests that every legal case is a battle: one side plays by the book, the other side tries to poke holes in it.

Even in strong democracies with "well-drafted" laws, loopholes still sneak in because the law is human-made, and humans are not perfect. No matter how well written the law is, someone will find a way to twist it. Loopholes are a by-product of a system built on interpretation, strategy, and endless dispute. Furthermore, when asked about the origin of loopholes, another lecturer argued that there is not just one section of the law that makes it prone to loopholes, but instead a bit of everything. He said: *"It's impossible that a piece of law drafted 164 years ago can cover every instance of offences against a person in 2025. ... Society develops and subsequent law has not been enacted or introduced to keep up with pace."*

He contends that loopholes are affected by everything, including poor drafting and a complex legal system, but oftentimes, they emerge because laws become outdated as society evolves. For example, the UK's Offences Against the Person Act 1861 is too old to fully address modern crimes. Another example is rapid developments, such as AI, creating new legal grey areas that existing laws cannot cover, making continuous reform essential.

5.3 Ethical Concerns with Loopholes

When people exploit legal loopholes, it raises complex ethical questions about intent, fairness, and professional responsibility. This subsection examines participants' views on whether loopholes are inherently unethical, the morality of exploiting them, and the justification of using a legal loophole.

When asked whose fault it is for loopholes being exploited, some participants placed the blame on the government and legal system, while others placed the blame on the user. One professor proposed that the fault lies with the state, saying: *"The responsibility should be on the state to make sure that law is properly drafted because there is an imbalance of power between the state and individual."* This participant largely placed responsibility on the government rather than the individual. He explains that in criminal law, the state holds the power to take away liberty, creating an inherent imbalance of power. While civil cases typically involve only financial loss, criminal proceedings carry far greater consequences, making it essential that laws are properly drafted and fit for purpose. As such, it is the responsibility of the legislature and the state to ensure the law functions as intended, and individuals should not be blamed for operating within the legal framework that governs them.

In contrast, a law student asserted that the fault lies in the hands of the user. He says: *"The top 1%, top 5% are the ones who use tax evasion schemes or tax planning schemes. So it's not that you're using these loopholes to save your skin, to save your soul. It's different from a roadside tea vendor or any street food vendor only accepting payment in cash because they can't afford to pay taxes. You can say that what they're doing might not be unethical, even though it's*

quite illegal.” He proposed that the ethicality of exploiting legal loopholes depends on context, intent, and consequences. Loopholes used for one's benefit are unethical because they deprive society of resources, while those used by the poor to survive may be less unethical. He further stresses that when laws are “patently evil or patently unjust,” exploiting loopholes can be morally justified.

In comparison, another professor stated that it does not lie in a specific party's hands but is the fault of everyone. He suggested that the government should develop a legal ethics infrastructure to prevent the abuse of loopholes; Lawyers should be cautious when encountering legal loopholes; Clients should also attempt to work with the law by asking their lawyers not to exploit legal loopholes.

When asked if loopholes are intrinsically ethical, most participants agreed that loopholes are not, stating other factors of loopholes to be more problematic than loopholes themselves. For instance, one professor said: *“It depends on the understanding and the purpose of the law and why it was introduced, and if the purpose of the law was not respected then yes.”* He argued the term's connotations often misrepresent the issue. He emphasised the role of judicial interpretation, noting that before 1911, judges could not consult parliamentary debates to clarify legislative intent. For example, in laws against “dangerous” driving, reviewing records of debates, such as concerns over a rise in speeding offences, can reveal the law's original purpose. If a loophole undermines that purpose, there is a stronger case for viewing its use as unjustified.

Moreover, another participant highlighted the importance of loopholes and their benefits to society, suggesting that loopholes could be ethical and beneficial. He stated: *“At times, they are quite productive as well. They allow us to deal with the changing social realities. They allow us to deal with the changing times, especially in trade and especially in technology. The law has to follow the merchant, not the other way around ... some of these loopholes allow us to reckon with what we are dealing with today.”* The participant asserted that loopholes are not inherently negative and can, in fact, be productive, particularly in helping the law adapt to changing social and economic realities, such as developments in trade and technology. However, a loophole only becomes problematic when it is misused in practice or when judicial intervention fails to correct abuse. Thus, the ethical dimension of loopholes lies less in their existence within legislation and more in how they are operationalised and addressed institutionally.

However, some participants emphasised the unethicity of loopholes through the intention to carry them out. One lawyer stated: *“It does allow individuals — particularly wealthy individuals — who can afford to engage the best lawyers and councils to help them exploit these loopholes.”* She illustrates that wealthy individuals are uniquely positioned to exploit tax loopholes, largely because they can afford to hire elite lawyers, accountants, and financial advisers who specialise in identifying and structuring around such gaps in the law. Abusing loopholes, especially those that are tax-related, is often legal yet ethically questionable, as they allow the ultra-wealthy to avoid contributing their fair share to public services while ordinary citizens, lacking access to such sophisticated tools, bear a disproportionate tax burden.

On top of that, another lawyer suggested that the usage of loopholes is never wrong. He emphasised: *“They haven’t broken the law by virtue of using this interpretation, there is nothing to say that they are wrong in doing so, other the law would say do not do that. So then, if legislators believe that is not the intention of the law, they would change that.”* He suggests a formalist approach to law. He frames loopholes as legally permissible unless explicitly prohibited. In this view, responsibility lies with legislators rather than individuals: if a particular interpretation is contrary to legislative intent, it is the duty of lawmakers to amend or clarify the statute. He further contends that creativity is an inherent feature of human reasoning and should not be constrained, even within the domain of legal practice.

Furthermore, when asked whether using loopholes could be justified if it was to bring justice to society, most participants claimed that in a perfect world, this could be possible; however, currently, this is not a viable solution. One participant suggested that *“It’s pretty much accepted now that Abraham Lincoln broke the constitution [because] he thought the law was wrong and breaking it was justified ... but you can only do that retrospectively.”* He suggested that while abusing loopholes can sometimes be justified when the law itself is perceived as unjust, this legitimacy can usually only be recognised retrospectively. Historical examples, such as figures who openly defied laws they believed were wrong, illustrate this notion. However, concerns were raised about the risks of providing individuals a licence to disregard laws whenever they choose, as this raises the critical question of who has the authority to decide when breaking the law is acceptable.

Additionally, another lawyer suggested that using a loophole to give justice to society was inherently wrong. He stated that everyone’s interpretation of ‘justice’ is different, leading to ambiguities regarding the action. This could lead to people committing unethical acts in the name of justice, undermining the written law.

5.4 Impacts of Loopholes on Wider Society

When asked about loopholes and their effects on society, many participants agreed that loopholes would break public trust. One professor said, *“Public trust can be repaired by updating laws and by making sure loopholes are addressed”* He acknowledged that loopholes are an inevitable feature of any legal system, but emphasised that the critical issue lies in how governments respond. When authorities publicly acknowledge loopholes and take steps to correct them, confidence in the legal system is strengthened. Public trust, they proposed, can be restored and maintained by updating legislation and ensuring loopholes are effectively addressed.

Contrarily, a professor expressed that even though in theory it should affect trust, that trust does not largely affect society; He says: *“Take the Panama Papers, for instance — it was a scandal, but ultimately, little changed.”* He noted that while the use of legal loopholes could harm public trust, there has rarely been a systemic collapse directly attributable to lawyers exploiting such gaps. The Panama Papers scandal certainly tarnished reputations, but yielded little structural change. He also suggested that Spain’s high-profile tax controversies involving Shakira and Messi have yielded little backlash and public defamation.

Lastly, when asked about which parts of law have the most loopholes, all participants stated that tax law was open to the most vulnerabilities.

5.5 Future Strategies to Prevent Loopholes

Participants were asked about strategies to prevent the use of loopholes. When asked who the responsibility of this job lies in, one participant stated: *"Law will always operate with language, and language will open to interpretation, but it's the judges and those who judge the law are the ones who must do that responsibly and reasonably"* He noted that eliminating loopholes is unrealistic, as laws always require context. Under the common law approach, judges and legislators continually refine meaning — such as determining what qualifies as “dangerous” driving — ensuring the law adapts to circumstances. The originalist approach instead grounds interpretation in the law’s original text, like the 1787 right to bear arms, but still requires modern contextualisation to keep legislation relevant. Both approaches rely on ongoing judicial and legislative interpretation to patch loopholes over time rather than achieving a once-and-for-all fix.

Participants were asked how AI could assist in patching and identifying loopholes. One law student discusses how AI could advance to have the ability to cover every scenario, stating: *"Maybe with AI, it can be so specific that there will be no contradiction at all. Every possible human action will have an already predetermined rule to judge whether it is right or wrong."* He discussed the potential role of AI in addressing loopholes, suggesting it could generate rules for an infinite number of possible situations by applying broad legal principles to specific scenarios. Unlike traditional law, which must remain abstract and general, AI could provide highly detailed, case-specific rules to reduce contradictions and prevent loopholes. However, concerns arose about accountability: while judges can be held responsible for wrongful decisions, AI cannot. As a result, the suggested role for judges would shift towards oversight, acting as appellate authorities to correct errors or absurd results produced by AI.

When participants were asked about the most important step in fixing issues with loopholes, one participant stated, *"Recognising that they happen is an initial step, when a loophole is exploited, we shouldn't be complacent."* He suggested that recognising that loopholes are an inevitable feature of any legal system is the most important step in fixing it. Authorities must first publicly acknowledge the existence of loopholes to be able to take steps to correct them, and the effects of loopholes can be suppressed by updating legislation and ensuring that loopholes are effectively addressed.

On the other hand, another participant leaned towards the role of judges and the need for a diverse judicial panel. He suggested: *"You would need judges who are insulated and protected from the overreach of the legislature and the executive, so they can render correct decisions, independent decisions. ... Judges should be appointed on the basis of merit and on the basis of diversity at the same time ... and appellate processes are very necessary. You cannot have a situation where whatever one judge says is all."* He indicated that loopholes cannot ever fully go away, but they could become smaller and less problematic. Firstly, he proposes that judges need

to be independent and not under the influence of politicians. Secondly, judges should be picked for knowing the law, but also reflect the diversity of the country so rulings make sense in the real world. Accessible appellate processes are also essential, since courts are “supreme but not infallible,” and errors must be open to correction. Ultimately, loopholes are patched through established interpretive procedures rather than permanently resolved through language.

Another two participants adopted a more practical view, focusing on how loopholes deprive marginalised communities of their basic needs. One suggests: *“Because these stakeholder consultations that should have happened initially or do not happen when they should have had or when the bills are being drafted, there is a huge gap between what the stakeholder actually faces and what the stakeholder actually would recognise as problems compared to what the bill is actually about. So I think a good first step would be to include different sorts of voices and predominantly include voices of the most important stakeholders for whom the bill is actually aimed at.”* The speaker emphasises that the first step in fixing loopholes is to centre the voices of primary stakeholders, particularly those most affected by legislation. She explains that many laws are drafted without meaningful consultation with the very communities they aim to protect. The failure to include marginalised perspectives will often result in conventions that look strong on paper but are disconnected from real-world needs, particularly in contexts like the Global South.

Lastly, a professor emphasised that the most effective way to address loopholes is by strengthening the legal ethics infrastructure. This involves creating clearer ethical rules, ensuring stronger enforcement, and improving legal education. Most importantly, they stressed the need to cultivate a professional ethos, where lawyers recognise that their duty goes beyond client advocacy; they also serve as representatives of the law itself, guiding clients to act responsibly within the legal system.

Overall, the interviews reveal that legal loopholes are seen as both inevitable and deeply contested. While some professionals regard them as intrinsic to the adversarial nature of law, others stress their potential to undermine justice, ethics, and public trust. Responsibility is both placed on governments and on individuals for exploitation. The impacts of loopholes range from eroding fairness and equality to occasionally serving as productive tools that allow the law to adapt to new social and technological realities. Most participants agreed that loopholes can never be fully eliminated, but they can be managed through better drafting, judicial responsibility, stakeholder inclusion, and, in the future, possibly AI. Ultimately, the findings highlight that loopholes are less about isolated flaws in the law and more about the ongoing problems between language, interpretation, and power within the legal system.

Conclusion

The issue of legal loopholes remains highly relevant and deeply concerning, as shown by the cases presented in this paper where their abuse has distorted justice. The objective of this paper has been to highlight how loopholes are not rare accidents but intrinsic features of the legal system, emerging repeatedly despite attempts to eliminate them. The findings confirm that

loopholes are frequent, systematically exploited, and profoundly problematic, eroding trust in the law and raising questions of fairness and accountability. While there is widespread recognition that loopholes exist and that their use is wrong, efforts to eradicate them have proven extremely difficult.

By situating these arguments within broader debates, this paper contributes to the wider research on legal loopholes, underlining the urgent need for further inquiry. Future research must explore not only technical reforms but also cultural and institutional changes that can reduce reliance on loopholes as tools of privilege. Until then, the persistence of loopholes will continue to threaten the legitimacy of legal institutions and the principles of justice they claim to uphold.

Works Cited

- ““Double Irish with a Dutch Sandwich””: Der Milliarden-Steuertrick "FINEXITY Blog.”
Finexity.com, 12 Jan. 2024,
finexity.com/en/blog/double-irish-with-a-dutch-sandwich-billions-tax-scheme.
- “Arguments for and against a Codified Constitution.”
Thepoliticsteacherorg.thepoliticsteacher.org,
thepoliticsteacherorg.thepoliticsteacher.org/arguments-for-and-against-a-codified-constitu
tion. Accessed 26 Feb. 2025.
- Bamigbola, Silas. “Legal Loopholes: Definition, Exploitation, and Real-World Examples.”
SuperMoney, edited by Andrew Latham and Ante Mazalin, 2024,
www.supermoney.com/encyclopedia/legal-loophole. Accessed 15 Mar. 2025.
- Baude, William, and Stephen E. Sachs. “THE LAW OF INTERPRETATION.” *Harvard Law Review*, vol. 130, no. 4, The Harvard Law Review Association, 2017, pp. 1079–147,
<https://doi.org/10.2307/44865509>. JSTOR.
- Bowers, Simon. “Google Pays €47m in Tax in Ireland on €22bn Sales Revenue.” *The Guardian*,
4 Nov. 2016,
[www.theguardian.com/business/2016/nov/04/google-pays-47m-euros-tax-ireland-22bn-e
uros-revenue](https://www.theguardian.com/business/2016/nov/04/google-pays-47m-euros-tax-ireland-22bn-euros-revenue). Accessed 11 Aug. 2025.
- Christie, George. Vagueness and Legal Language. *Minnesota Law Review*, 1964, pp. 885–911,
[scholarship.law.duke.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1207&cont
ext=faculty_scholarship](https://scholarship.law.duke.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1207&context=faculty_scholarship). Accessed 21 Feb. 2025.
- Cornell Law School. “Legal Realism.” LII / Legal Information Institute, June 2023,
www.law.cornell.edu/wex/legal_realism. Accessed 1 Sept. 2025.
- Encyclopædia Britannica's Editors. “Napoleonic Code | Definition, Facts, & Significance.”
Encyclopædia Britannica, 19 Jan. 2018, www.britannica.com/topic/Napoleonic-Code.
Accessed 21 Feb. 2025.
- Endicott, Timothy. “Law and Language.” *Stanford Encyclopedia of Philosophy*, edited by
Edward N. Zalta, Spring 2022, Metaphysics Research Lab, Stanford University, 2022,
plato.stanford.edu/entries/law-language/#UseLangLaw.
- Haag, Matthew. “Man Who Lived Rent-Free in New Yorker Hotel Is Found Unfit to Stand
Trial.” *Nytimes.com*, *The New York Times*, 6 Nov. 2024,
www.nytimes.com/2024/11/06/nyregion/mickey-barreto-unfit-trial.html. Accessed 3 Apr. 2025.
- Hart, H. L. A. *The Concept of Law*. 1961. Edited by Joseph Raz and Penelope A. Bulloch, 3rd
ed., Oxford University Press, 2012.
- Hayes, Adam. “Loophole.” *Investopedia*, 2019, www.investopedia.com/terms/l/loophole.asp.
- Hopkins Centrich PLLC. “Always Show up in Court Even If It Seems Frivolous.” *Hopkins
Centrich Law*, 6 Apr. 2024,
shareholderoppression.com/blog/always-show-up-in-court-even-if-it-seems-frivolous.
Accessed 3 Apr. 2025.

- Jong-Fast, Molly. "GOP Digs up 'Zombie Laws.'" *Vanity Fair*, 18 Dec. 2024, www.vanityfair.com/news/story/republicans-zombie-laws-comstock-act-abortion. Accessed 15 Mar. 2025.
- Katz, Leo. "A Theory of Loopholes." *The Journal of Legal Studies*, vol. 39, no. 1, Jan. 2010, pp. 1–31, <https://doi.org/10.1086/649046>.
- Magazine, Smithsonian, and Kate McMahon. "Stressed about Taxes? Blame the Ancient Egyptians." *Smithsonian Magazine*, 3 Apr. 2024, www.smithsonianmag.com/history/stressed-about-taxes-blame-the-ancient-egyptians-180984059/.
- Murphy, Chris B. "How People Use the Double Irish with a Dutch Sandwich to Reduce Taxes." *Investopedia*, 2019, www.investopedia.com/terms/d/double-irish-with-a-dutch-sandwich.asp.
- Nelson, Olivia. "Unveiling Legal Loopholes: Examining Controversial Legal Tactics – Legal Eaglefirm." *Legaleaglefirm.uk*, 31 Mar. 2024, legaleaglefirm.uk/unveiling-legal-loopholes-examining-controversial-legal-tactics.com. Accessed 5 Feb. 2025.
- Onion, Amanda, et al. "Napoleonic Code Approved in France." *History*, A&E Television Networks, 28 July 2019, www.history.com/this-day-in-history/napoleonic-code-approved-in-france. Accessed 21 Feb. 2025.
- Organ, Paul R., and Navodhya Samarakoon. "Doubling down on Tax Avoidance: The Effect of Double Irish Closure on the Profit Shifting of U.S. Multinational Companies." *International Tax and Public Finance*, Springer Science+Business Media, Apr. 2025, <https://doi.org/10.1007/s10797-025-09888-7>. Accessed 23 Apr. 2025.
- Palma, Stefania. "Judge Blocks Trump's Move to Use Wartime Law to Deport Venezuelans." *FinancialTimes*, *Financial Times*, 16 Mar. 2025, www.ft.com/content/7ecc942c-d7d8-4eb4-a231-52ef729de110. Accessed 15 Mar. 2025.
- Posner, Richard. "Case Western Reserve Law Review Legal Formalism, Legal Realism, and the Interpretation of Statutes and the Constitution Recommended Citation." 37 *Case W. Res. L. Rev.*, vol. 37, no. 2, 1986, case.edu/law/sites/default/files/2019-08/Legal%20Formalism%20Legal%20Realism%20and%20the%20Interpretation%20of%20Statute.pdf.
- Sauerborn, Philipp M. "Update: Double Irish with a Dutch Sandwich—How Apple and Others Slash Their Tax Bills - Philipp Sauerborn's Blog." *Philipp Sauerborn's Blog*, 29 Mar. 2017, philippsauerborn.com/en/update-double-irish-with-a-dutch-sandwich-how-apple-and-others-slash-their-tax-bills/. Accessed 10 Sept. 2025.
- Schneiderman, Eric. *TENANTS' RIGHTS GUIDE* Attorney General. www.nyc.gov/assets/buildings/pdf/tenants_rights.pdf.
- Víctor García Yzaguirre, and Hugo Osorio Morales. *Restrictive Interpretation of Exceptions: Identification and Hunt of Ideologies*. no. 4, Nov. 2023, pp. 721–44, hrcak.srce.hr/file/449427. Accessed 21 Feb. 2025.

The Fascinating Use of AI in Marketing By Ishaan Abhyankar

1. Introduction

Artificial intelligence has advanced rapidly, transforming the business world and becoming a useful tool for promotion, marketing, and data analysis. Typically used for human-assigned tasks, AI uses machine learning and algorithms to understand data and optimize business operations. AI's global impact has been profound, opening new pathways for interacting with customers and marketing methods. (Al-Gahmdy, 2021)

Within marketing, AI has been used in many ways to enhance the quality of understanding consumers, marketing and advertising, and supply chain management. In understanding consumers, AI assists companies in tracking purchase and online patterns of consumers, predicting future behaviors, and giving personalized recommendations based on this data. This data helps firms develop marketing strategies catered to the data that AI attains and interprets. In Advertising and promotion, AI is used to generate advertisements, target audiences with accuracy, and improve costs and efficiency. However, with AI content creation comes many ethical debates from the consumer end. (Brei, V.A. (2020).

Marketing Analytics represents another avenue where AI is seeping into the business world and changing decision-making. By processing various amounts of data, AI identifies market trends and forecasts demand at an exceptional rate. Companies are not only using AI for interpreting data but predicting future decisions in consumers.

Despite its rapid adoption, existing literature lacks real-world cases of AI being effective and failing in the business world. It also does not address the debates in which consumers express their opinions on AI being used in companies. Most previous research focuses on what AI is and how it is used, rather than the social approach or the real effects on companies and the consumer. This paper addresses the gap in real-world application of AI in the business world by diving deep into companies like Mango, L'Oréal, and H&M. This study aims to explore how AI improves efficiency while being the catalyst for ethical questions and debates.

2. Literature Review

2.1 AI's application in understanding consumers

Marketing's primary goal is to understand the consumer's needs and behaviors to design, advertise, and deliver products that provide value (AMA, 2019; Lakshmipriyanka et al., 2023). A crucial aspect of marketing is the value proposition and delivery to customers, including goods and services, as well as any other necessary elements to fulfill the customer's needs. The definition of marketing (AMA, 2019). This involves optimizing marketing campaigns and enhancing customer experiences to purchase a product (Lakshmipriyanka et al, 2023).

AI in marketing can be seen as a controversial topic due to the ethical concerns of firms that utilize it. Many forms of AI can gather, source, and analyze personal information such as browsing histories and purchasing behaviors (Stamatogiannakis et al., 2019).

Personalization can exploit psychological preferences and subtly persuade consumers toward specific products. Also, some AIs take advantage of customers based on their psychological preferences. Often, AI can persuade a customer to purchase a certain company's product based on its targeted advertisements (German et al., 2019)

AI application also boosts speed and efficiency within the marketing industry. Platforms like Facebook have AI-generated advertisements and target these ads to various customers through content the user has already uploaded, producing content much faster than human intelligence (Maria Christina, 2020). AI can also be used for promotion, such as competitive-based pricing and incentives, such as gifts and rewards.

There are instances where AI is utilized, such as specialized offers and discounts for planned advertisement campaigns. Other examples include offering discounts, competitor-based pricing, and location-based promotion (Bieliaiev, 2022).

Beyond the purchase, AI is a keystone to developing customer-brand relationships by enhancing efficiency with shopping, such as shopping and personalization post-purchase, such as gifts and rewards, to keep this relationship strong (Mazurek G, 2019). Increasingly, AI advertising is becoming more relevant because customers trust apps to collect data from them and serve them better. Most advertisements will reside on an app and not pop up when the customer is not expecting it, to limit the intrusiveness of AI advertising (Li, 2019).

Despite many advantages, AI has often struggled with emotional intelligence, a critical marketing component. To address this, chatbots and voice assistance were introduced into the business world to deliver a relevant, human-like, and emotional answer. 2021. In addition, there is a wide variety of ethical concerns with organizing and analyzing data. Data protection is a cornerstone of ethical marketing practices, primarily as AI technologies increasingly collect and analyze consumer data.

Overall, AI has a massive significance in digital marketing as it is used to understand and organize customer behavior and data. This leads to the production of targeted digital advertisements and to the detection of patterns in preferences.

2.2 AI's application in advertising and promotion

The rapid development of the internet is transforming traditional forms of advertising and product promotion. Firms are utilizing AI to analyze data and match user needs, thereby creating intelligent advertisements. (Sumani, 2022).

For example, we see AI being used in many forms of promotional content creation, such as social media posts, text messages, presentations, and digital advertisements. Such use of AI also improves financial situations, with firms spending less capital on showing ads to customers who are not in their target audience. (Uchakova, 2022).

Similarly, AI can create advertisements and promote based on data it receives from consumers, which it can analyze for consumer behavior, and effectively create targeted advertisements.(Deng et al. 2019). It is important to note that AI promotion and advertisement significantly lowers the cost of creation, since employment costs are negated.

The purpose of creating advertisements is to promote your product in a way that will sell your value proposition to the customer and gain relevant feedback on your product. Using AI ads, companies can organize their customers' wants and needs and help them take the next step in considering purchasing the product. (Arora, 2024).

While there are apparent benefits to AI-driven advertisements and promotion, it comes with an ethical dilemma. For example, when a survey was conducted on the reason consumers use ad blockers, the most notable reason was invasion of privacy, which is what AI does when it collects and analyzes consumer data on their purchases, searches, and more. (Naz, 2024).

This feedback from consumers raises various concerns for firms on how much they should use AI advertising and whether or not they think it will be effective. Most notably, how will consumers react to these AI advertisements? Is there a fear of AI? (Boudri, 2025). These questions are fundamental for companies to consider, as consumer reactions to AI can directly correlate to the financial forecasts and sales.

2.3 AI's application in marketing analytics

Marketing analytics, a data-intensive function, is the idea and execution of gathering and analyzing data from various marketing strategies to understand performance and track consumer behavior. Generative AI is being used in most companies to enhance the understanding of data through generative algorithms that can interpret the data.

In marketing analytics, there are many ways AI is being implemented to carry out specific functions.

Expert systems can be considered computer-generated systems that improve some aspect of human nature and decision-making through pattern recognition in data. Companies have used Tableau, a marketing analysis programming tool incorporating AI, to do this. (Mortimer, 2018) Similarly, AI is used by companies for instances such as when and why a customer feels the need to stop buying a particular product or service. AI can detect early signs of customer dissatisfaction so the company can adapt its marketing plan accordingly to attract customers. After all, business is all about the customer and selling a unique story.

The most notable way for AI to be used in marketing analytics is to stay ahead of the market and analyze trends. AI forecasts trends, identifies potential customer actions, and optimizes marketing strategies.

AI can delve into the decision-making and attitudes of customer data, categorizing consumers by personality traits, values, interests, and lifestyles, which can be incredibly useful for crafting personalized messaging and offers (Russell, 2025).

Also, transparency in AI marketing tools refers to the clarity and openness with which companies disclose the workings of their AI systems, including how data is collected, processed, and used in decision-making processes (Russel,2025).

Existing literature lacks real-life examples of how AI is being used in companies nowadays. Our paper will focus on examining the struggles and successes of AI in the marketing and business world, including the benefits of using AI in business, customer opinions on AI, and the downsides of using AI as it pertains to privacy and ethics.

However, although existing literature provides the foundation for AI use in business, it lacks the application of these ideas to the real world. The case study method is a reliable and functional way of conducting research, as previous literature lacks the incorporation of real-world examples that relate directly to AI in the business world, specifically in advertising and promotion. Unlike experimental methods, the case studies used focus on how cultural and organizational actors lead to the use of AI in marketing, such as a certain company's goals in doing so.

Information was searched in a variety of ways. One being the discovery of companies that have used AI, for the better or worse of their company, and how it has affected the general public. Information was directed in this paper to aim at not only the methods of AI used in business, but also the pros and cons of doing so using customer feedback.

3. Methods

This study uses the case study approach to examine how Artificial Intelligence (AI) is used within the marketing field. The case study method was chosen because it provides a unique vantage point from consumers and the company itself on the positives and negatives of using AI in the business world. Case studies take into account the cultural and operational factors of AI use, something surveys or other forms of research won't address.

The research focuses on 3 companies: Mango, L'Oréal, and H&M. These companies were chosen due to their use of AI in various fields of marketing, with sufficient data and consumer input to back it up. These 3 companies provide avenues to compare and contrast AI use, ethics, and operational decisions within the company. Mango emphasizes how AI was used in advertising and promotion, L'Oréal emphasizes AI-powered tools for personalization, and H&M explores operational efficiency in forecasting demand.

Information was obtained from secondary sources, including journal articles, articles listed on each company's website, and press releases. Additional consumer feedback was analyzed from online platforms such as TikTok and Reddit. These methods allow for a two-fold vantage point, one of the firm and the other of the consumer.

4. Case studies

4.1 Mango

Mango is a fashion retailer located in Spain, known for its design, production, and the sale of clothing and various accessories, ranging from jewelry to handbags. Mango has over 16,000 employees and over 223 stores in 115 markets, making it a massive global company.

The first instance where Mango used AI was for their first campaign, Teen Line Limited edition, where they used generative AI trained to position Mango's clothing pieces on a model to enhance the quality of images and market their product. (Rimmer, 2025). After this, the art team at Mango edited and touched up the photos slightly to mold the images into an advertisement campaign. Even the models used for these images were AI-generated, not just the clothing and the backgrounds. After the customer buys a particular product, AI-generated recommendation algorithms reach customers and provide next steps, including similar designs and the pricing policy on [mango.com](https://www.mango.com). This advancement was part of Mango's mission to expand through technology and operational excellence. (Jin, 2024).



Figure 1: AI-generated image for Mango's Teen Line, Limited edition

In this image, Mango used generative AI to put the clothing onto the model for content creation.

Mango received positive feedback from customers and the general public about their first implementation of AI. Olivia Campbell argued, "If the imagery looks good, feels premium, and sells the product without distracting errors, that is all that matters." (Campbell, 2025).

On the other hand, they also received negative feedback for using AI in modeling. The images and models created were deceiving and might not be the product itself in real life. (Nat, 2024). Mango did not explicitly say that their images were AI-driven, leading to public debate. Some may argue that this violates privacy because Mango collects customers' data and uses it to their advantage. On the other hand, customers might not want recommendations while shopping because some customers consider it a distraction. (Reddit, 2026)

The second implementation of AI in Mango's development in content creation was to reduce production prices. They replaced real-life models with AI-generated models because it was significantly cheaper than human counterparts. (Lilyotron, 2024). According to the CEO of

Mango, Toni Rooz, it is about faster content creation. (Snyder, 2018). Brandnat on TikTok argued that using AI was cost-efficient, unlike Mango flying out their team to a particular site, including photographers. Mango said that they were using AI solely as a co-pilot and to broaden the creativity of their employees. Sofia Luzón Atelier argued against what Mango said about AI being a co-pilot, and instead argued that they used AI to complete the task for them. In addition, Mango's profits from digital advertising quadrupled.

4.2 L'Oréal

The L'Oréal story begins in 1909, when Eugène Schueller, a chemist with a vision, started by producing hair dyes and then transformed it into a clothing company. They have over 94,000 employees and over 37 factories worldwide, suggesting they are one of the largest companies in the fashion industry. AI is becoming extremely popular in the fashion industry, and it is used for many occasions, including modeling and personalization (L'Oreal, 2015).

L'Oréal has been using AI in many ways. For example, the development of L'Oréal Paris' Skin Genius is an AI-powered tool that customers can use to understand their skin's specific needs, which is followed by personalized recommendations of products that will solve these issues. The AI-powered tool has also been used in research for skin aging, including factors such as sunlight exposure (Shong, 2025).

The reason for developing Skingenius was that there are various skin types and skin factors, such as color and resistance. So, L'Oréal argued that skincare should be customized to one's face rather than having a generic solution.

To meet these goals, there are three steps to using SkinGenius. First, the customer takes a selfie of their face and uploads it. Then, the customer puts in their personal information, including age and skin type. Finally, an AI-backed platform calculates the results and gives the customer feedback on the eight attributes of their face. To protect privacy, L'Oreal does not store the selfie; instead, it deletes it (L'Oreal, 2015).



Figure 2: L'Oreal shows how Skin Genius is correctly used on their website

In this image, L'Oréal shows how to correctly use Skin Genius, with the first step being a picture of the customer's face without makeup and glasses. As you can see in the image, L'Oréal gives constructive feedback on the customer's face.

However, public opinion on SkinGenius is ambiguous because there are no reports of benefits or customer feedback related to it. There is no clear answer as to whether or not Skin Genius was indeed used by customers, and what their opinions and reactions were. Real feedback does not exist on Skin Genius, so there are no measures of impact on either L'Oréal or the consumer.

L'Oréal not only used AI for its website tools to enhance customer experience, but it also used it in content creation. Using Google Ads' 8-second Veo, L'Oréal was able to turn AI-powered images into 8-second ads that were put on display to the public. The reason for this was cost efficacy- in 2023, over 60% of the company's digital media budget went into social platforms.

However, AI ads by L'Oréal were deemed weird and inappropriate when they were watched. According to Reddit, an AI-powered commercial was played in which the AI voice from L'Oréal mispronounced a few words, making it apparent it was created by AI. Many of the following comments on this Reddit post had phrases such as "that is very weird" and "Wtf." Many were not happy with the use of AI and were caught off guard.

Finally, L'Oréal was having a difficult time understanding the effect of its advertising money on sales and results across 20 brands and 4 Nordic countries. So, they needed a way to track which campaigns worked in terms of positive results and which campaigns did not.

Their solution was a custom AI tool called Tidal, which aimed to predict campaign outcomes in e-commerce before they were launched by using data. (Misra, 2025). It was also used to cut down wasted capital on advertisements that were not doing as well, and use that money to enhance advertisements that were doing well.

The effects of Tidal were more than apparent, with a 22% in media and brand efficiency, 14% better performance when shifting investments, and 8% increase in overall efficiency (Zwieglinska, 2025).

4.3 H&M

H&M is a Swedish multinational fast fashion retailer headquartered in Stockholm. Known for its fast fashion business model, H&M sells clothing, accessories, and homeware. The company has grown to one of the most popular fashion stores across many different age groups due to its affordable and trendy clothing designs and options.

H&M's earliest adoption of AI focused on demand prediction and inventory management, a critical area for fast fashion retailers. Like many retailers, H&M faces numerous challenges with storing inventory, as well as the unpredictability of consumers' preferences and their decision-making. So, machine learning and advanced analytics through AI were used to detect and understand seasonal trends within their customer base and market trends to forecast

sales and revenue. The company invested a portion of its capital into a strong data interpretation system that analyzed sales transactions and consumer reactions, giving AI data to work with. Then, AI was implemented to make automatic decisions based on inventory allocation and space. (DigitalDefynd, 2025).

The benefits of using AI to predict consumer decisions, seasonal trends, and inventory allocation were apparent. The company achieved a significant reduction in overstocking, which can play a critical role in the financial health of the company. Due to this, customers were more satisfied because there was enough supply to meet their demands.

Secondly, H&M for many years had a problem with high return rates for customers, mostly due to sizing and order discrepancies. Customers could not identify their size correctly and were seemingly frustrated. Having a high return rate on items was hurting customer satisfaction as well as environmental concerns. So, H&M turned to AI-powered 3D fitting rooms to assist customers with their fitting. Also, they created a platform where customers could input their body type and measurements using their camera, which AI would then create a persona of their body. AI would then create an avatar platform where customers could see the article of clothing on their body type. This implementation boosted customer satisfaction and lowered return rates, and was an overall success.

5. Discussion and Implications

AI's application in the business world is a polarizing topic, as there are many fields where AI is benefiting customers and businesses, including marketing analytics. This is AI's ability to recognize and understand market trends, operations, and predictions. AI in marketing analytics relies on the quality of data, which is far less of a challenge for firms because they don't have to use humans or resources to complete tasks. For example, AI can simply predict market trends such as spikes, enabling firms to get ahead of the trends in the market, as well as improving efficiency and producing less waste. This domain is less challenging because it is largely data-driven. As long as companies have access to reliable datasets, the results are trustworthy and beneficial.

Secondly, advertisements and promotions have a more neutral position. AI allows companies to reduce costs by spending fewer resources on humans who complete tasks, so the money can go into other departments of the company. Companies such as L'Oreal and Mango have used AI to cut costs, accelerate content creation, and improve efficiency within the company. However, the use of AI in marketing and advertising can lead to ethical and perceptual concerns. AI-generated advertisements may appear inauthentic or fake because the figures in the advertisements are AI-generated, so customers don't know how the product looks or is being used, as seen before in the image created by Mango.

The most challenging part of AI in business remains in understanding consumers and their behavior on an emotional and ethical level. By using AI in this category, it is mostly reliant on invading the privacy of consumers and tracking their behaviors through data. While tools like L'Oréal's SkinGenius aim to personalize recommendations, their effectiveness depends on trust

and consent. AI has a challenging time replicating emotional intelligence, and there are many ethical concerns surrounding privacy and manipulation.

For Mango, an improvisation can be made with the way they used AI in their advertisements. Instead of letting AI make the images for the advertisements, they could have AI give them suggestions on what to incorporate in their advertisements and which audience to target. This might not be the most trustworthy method, but it will waive off ethical concerns that customers have.

6. Conclusion and Future Research

To summarize, AI analytics is the most effective in data-driven applications such as forecasting and inventory management. It provides benefits such as predicting and understanding market trends, reducing waste, and boosting overall efficiency. Companies such as Mango and L'Oreal used AI to cut costs, accelerate the process of content creation, and scale. While the use of AI is beneficial for companies financially, AI raises ethical concerns and authenticity debates on whether using it is effective. With personalization, L'Oreal's Skin Genius shows the potential in AI to create personalization for customers. The overall implications are mostly positive, with AI excelling in efficiency driving and data-heavy tasks such as supply-chain management, inventory management, and financial forecasting. However, from a consumer standpoint, AI can't be trusted due to ethical concerns, such as AI tracking consumer behavior on websites.

Future research should be focused on customer perception and how trustworthy consumers really feel about AI. Understanding this can benefit companies, as not only can they create AI personalization models, but they can also make ones that are accepted by customers. From a financial standpoint, future research can be directed towards the impact of AI, such as profitability, and long-term brand reputation. There is not a lot of research and evidence with real cases about the use of AI in business.

Works Cited

- Arora, N., Gandotra, R., & Dhiman, N. (2024). Consumer happiness and sustainable consumption. *Global Business and Organizational Excellence*, 43(6), 59–73.
<https://doi.org/10.1002/joe.22258>
- Boudri, R., Boudri, I., & Bentalha, B. (2025). Consumer perception of AI-Generated content in influencer campaigns. In *Advances in marketing, customer relationship management, and e-services book series* (pp. 187–204). <https://doi.org/10.4018/979-8-3693-3799-8.ch010>
- Brei, V. A. (2020). Machine Learning in Marketing: overview, learning strategies, applications, and future developments. *Foundations and Trends® in Marketing*, 14(3), 173–236.
<https://doi.org/10.1561/17000000065>
- Burnaman, E. (n.d.). Human-Powered Marketing in the Age of AI: How Snyder Group Blends Responsible AI with Authentic Creativity.
<https://resources.snydergroupinc.com/groupthink/human-powered-marketing-in-the-age-of-ai-how-snyder-group-blends-responsible-ai-with-authentic-creativity>
- Deng, S., Zhao, H., Fang, W., Yin, J., Dustdar, S., & Zomaya, A. Y. (2020). Edge Intelligence: The confluence of edge computing and artificial intelligence. *IEEE Internet of Things Journal*, 7(8), 7457–7469. <https://doi.org/10.1109/jiot.2020.2984887>
- DigitalDefynd, T. (2025, June 30). 5 ways H&M is using AI [Case Study] [2025] - DigitalDefynd. DigitalDefynd. <https://digitaldefynd.com/IQ/hm-using-ai-case-study/>
- Enache, M. C. (2020). AI for Advertising. *Annals of Dunarea De Jos University of Galati Fascicle I Economics and Applied Informatics*, 26(1), 28–32.
<https://doi.org/10.35219/eai1584040978>
- Haleem, A., Javaid, M., Qadri, M. A., Singh, R. P., & Suman, R. (2022). Artificial intelligence (AI) applications for marketing: A literature-based study. *International Journal of Intelligent Networks*, 3, 119–132. <https://doi.org/10.1016/j.ijin.2022.08.005>
- Hoelscher, J., & Mortimer, A. (2018). Using Tableau to visualize data and drive decision-making. *Journal of Accounting Education*, 44, 49–59.
<https://doi.org/10.1016/j.jaccedu.2018.05.002>
- L'Oréal. (2023, January 16). L'Oréal at CES 2024. L'Oréal.
<https://www.loreal.com/en/group/culture-and-heritage/l-oreal-history/>
- L'Oreal Case Study: AI attribution | Choreograph. (n.d.).
<https://www.choreograph.com/resource/loreal-super-charges-paid-social-efficiency-by-22-percent>
- Lparsons. (2025, April 14). AI Will Shape the Future of Marketing - Professional & Executive Development | Harvard DCE. Professional & Executive Development | Harvard DCE.
<https://professional.dce.harvard.edu/blog/ai-will-shape-the-future-of-marketing/>
- Mango creates the first campaign generated by artificial intelligence for its Teen line. (n.d.). Mango Fashion Group.
<https://www.mangofashiongroup.com/en/w/mango-crea-la-primera-campa%C3%B1a-generada-con-inteligencia-artificial-para-su-l%C3%ADnea-teen>
- MANGO's AI models: Empowering creativity or reinforcing outdated ideals? (2024, November 18). Lily Snyder.

- <https://lilyotron.blog/2024/11/18/mangos-ai-models-empowering-creativity-or-reinforcing-outdated-ideals/>
- Naz, H., & Kashif, M. (2024). Artificial intelligence and predictive marketing: an ethical framework from managers' perspective. *Spanish Journal of Marketing - ESIC*. <https://doi.org/10.1108/sjme-06-2023-0154>
- Priyanka, A. L., Harihararao, M., Prasanna, M., & Deepika, Y. (2023). A study on Artificial intelligence in Marketing. *International Journal for Multidisciplinary Research*, 5(3). <https://doi.org/10.36948/ijfmr.2023.v05i03.3789>
- Skin Genius Personalized Skin Analysis Tool - L'Oréal Paris. (n.d.). <https://www.lorealparisusa.com/skin-genius-landing-page>
- Sova, O., Bieliaieva, N., Antypenko, N., & Drozd, N. (2023). Impact of artificial intelligence and digital HRM on resource consumption within a sustainable development perspective. *E3S Web of Conferences*, 408, 01006. <https://doi.org/10.1051/e3sconf/202340801006>
- Stafford, M. R., & Faber, R. J. (2005). Advertising, promotion, and new media. *Choice Reviews Online*, 42(08), 42–4731. <https://doi.org/10.5860/choice.42-4731>
- Ushakova, I., Hrabovskyi, Y., & Szymczyk, K. (2023). ANALYTICAL DECISION LIST ALGORITHM FOR MANAGING CUSTOMER REACTION TO A MARKETING CAMPAIGN. *Polish Journal of Management Studies*, 28(1), 360–379. <https://doi.org/10.17512/pjms.2023.28.1.21>
- Varian, H. R. (2014). Big Data: New tricks for econometrics. *The Journal of Economic Perspectives*, 28(2), 3–28. <https://doi.org/10.1257/jep.28.2.3>

Do Behavioral Biases Hit Harder in Crypto? A Comparative Literature Review of Behavioral Biases in Crypto and Stock Markets By Aydinalp Erdeve

Abstract

Although behavioral biases are central to financial economics, their effects differ across asset classes. This research compares three core biases-overconfidence, herding, and attention-driven trading between equities and cryptocurrencies. This review documents how overconfidence increases turnover and leverage losses, how herding shapes collective moves, and how attention shocks create cycles of hype and collapse. In this paper, I will provide evidence showing that while these forces appear in stocks, they are amplified in crypto due to 24/7 trading, retail dominance, low liquidity, and social media. This research highlights the key differences and similarities between equities and crypto, by bringing up existing literature to evaluate whether cryptocurrency markets are more prone to behavioral biases than traditional stock markets.

Key Words Behavioral Finance, Overconfidence, Herding, Investor Attention, Cryptocurrencies, Stock Market, Fear of Missing Out (FOMO)

1. Introduction

Behavioral finance shows that investors often move away from perfect rationality. Biases such as overconfidence, herding, and attention-driven behavior regularly shape how people trade and how markets evolve (Hirshleifer, 2015). These forces come from both individual psychology and from social interaction, but the way they affect prices depends heavily on the design of the market. Who trades, what rules exist, and how information spreads all matter. Cryptocurrency markets differ from traditional stock markets in many of these aspects. They operate all day and all week, they have a high share of retail investors, and they are deeply connected to online platforms and social media. Because of these features, a key question arises: Are cryptocurrency markets more prone to behavioral biases such as herd behavior, overconfidence, and fear of missing out than traditional stock markets?

The stock market is a great place to start looking for proof. Research indicates that overconfidence leads investors to engage in excessive trading, miscalculate risks, and sustain exaggerated beliefs of their own proficiency (Odean, 1999). This kind of trading often leads to lower returns and increased volatility. Recent studies have identified similar phenomena in cryptocurrencies, where historical returns result in heightened trading volumes indicative of overconfidence, often surpassing those observed in equities (Tran, 2021). When investors disagree and short selling is hard to do, prices tend to go up since optimists are more likely to buy. This can also lead to speculative bubbles. There is already a lot of knowledge about these kinds of dynamics in stocks, but they may be more stronger in crypto, where there is more disagreement, markets are always open, and people who are pessimistic often have fewer ways to say what they think.

An additional trend that financial assessments effectively elucidate is herding. It elucidates the extent to which investors emulate one another's actions, provided that information sharing can account for it. Philippas et al. (2020) note that herding across cryptocurrencies can be even more pronounced than in traditional financial markets, particularly during periods of volatility. Herding can result in significant booms and profound crashes, as evidenced by historical events such as it happened in pandemics and bankruptcies. Cryptography simplifies herding significantly. Entry fees are low, markets never shut, and online narratives, which are often brief and emotional, spread rapidly. As a result of these social dynamics, cryptocurrency markets may experience more frequent and severe herding than equities.

The information environment adds another layer to the distinction between cryptocurrencies and stock markets. Research on stocks has shown that news stories, media attention, and search behavior can all move asset prices. Reviews of textual analysis and sentiment show that attention itself can be a tradable signal. In crypto, this effect is magnified because information flows almost entirely through social platforms like Twitter, Reddit, or Telegram. These channels encourage rapid attention cycles, leading to what is often called the fear of missing out (FOMO): when a sudden rise in attention causes many new investors to enter quickly, pushing up trading volumes and short-term prices.

Investor type also matters. Reviews comparing retail and institutional traders show that retail investors are more prone to biases and attention-driven trades, while institutions are steadier. Shen, Urquhart, and Wang (2019) say that because small buyers make up most of cryptocurrency markets, behavioral biases have a larger effect than in stock markets.

These biases are affected by the way bitcoin markets are set up. Biases can grow in a market where shortselling is limited, trading platforms don't work and there are high-leverage options. Copy-trading may be easier if the blockchain is open, and the constant circulation of money through stablecoins may make sudden inclines in demand worse. Instead of creating new biases, the way cryptocurrency is structured can make biases that already present in stock markets more powerful.

In this study, I will build on these fundamental insights by directly comparing the intensity of behavioral biases in cryptocurrency markets versus traditional stock markets. The focus will be on three core biases: overconfidence, herding, and attention-driven trading behavior such as the "Fear of Missing Out". By looking at previous literature, I will first review how these biases are present in stock markets and then examine whether similar or stronger patterns appear in crypto markets. Particular attention will be given to the role of retail investors, the influence of online platforms, and institutional market features that may increase the magnitude of behavioral effects. By bringing evidence from both asset classes, the aim is to evaluate whether cryptocurrencies are vulnerable to behavioral biases or whether they simply reflect established patterns in a new and different market design. This analysis will provide a clearer understanding of how market design and investor types shape the presence of some biases, and what this means for volatility and efficiency of the asset. This literature review compares how these biases appear differently in cryptocurrencies and traditional stock markets.

2. Conceptual Foundations

2.1 Theoretical Foundations

Behavioral finance was created as a response to the limitations of the Efficient Market Hypothesis (Fama, 1970), which assumes that markets are completely rational and fully efficient. This theory suggests that all the investors in a market are rational-decision makers. Researchers began to observe the differences in stock and cryptocurrency pricing and trading that could not be explained by traditional market models. Daniel, Hirshleifer, and Subrahmanyam (1998) developed formal models showing how overconfidence and biased self-attribution can drive excessive trading, momentum in returns, and long-term reversals. Hirshleifer (2015) further integrated findings across psychology and finance, showing that limits to attention constraints, and sentiment all create predictable departures from efficiency. These theories also underline the role of disagreement and short-sale restrictions: when pessimists cannot easily express their views, optimistic investors who consistently are buying an asset without thinking of the consequences, dominate price setting, which can lead to high increase in prices and higher volatility.

Alongside these researches, theories of social learning and information floods (Banerjee 1992; Bikhchandani, Hirshleifer, & Welch, 1992) explain how investors imitate others even when their own private information suggests otherwise. If a few well-known investors act in one direction, the other individual investors may follow them, since they are afraid of missing some crucial signals and earning money opportunities. Such cascades help explain why herding is common in markets, especially during uncertain times. Together, these theoretical foundations provide a strong basis for applying behavioral finance concepts to both traditional and emerging markets, such as it happens in cryptocurrencies.

2.2 Traditional Stock Markets

In stock markets, behavioral biases are well documented. Barber and Odean (2001) found that individual investors trade excessively, lowering their returns compared to market benchmarks which is a clear sign of overconfidence. Apergis et al. (2022) extended this evidence to G7 countries, showing that investor overconfidence significantly explains abnormal returns and trading activity. Christie and Huang (1995) and Chang, Cheng, and Khorana (2000) provided strong evidence of herding by showing that stock return distribution narrows in periods of market stress, indicating investors are moving in the same direction beyond what fundamentals justify.

Attention and psychology also influence the actions of an investor. Da, Engelberg, and Gao (2011) introduced Google search volume as a direct measure of investor attention and showed it predicts abnormal stock returns and volume. Other studies show that media coverage (Tetlock, 2007) and earnings announcements (Hirshleifer et al. 2009) cause short-term overreactions followed by corrections. These findings illustrate that even in well-regulated and institutionally diverse markets, psychological and social forces affect trading patterns and price dynamics.

2.3 Cryptocurrency Markets

There is a distinct environment that is presented by cryptocurrency markets, where structural characteristics may have an effect on behavioral biases. On the other hand, unlike equities, cryptocurrency is constantly trading, and there are relatively few possibilities for short selling. Logically, participation is dominated by retail investors, which are individuals who are not professionals and who purchase and sell stocks for their own personal accounts. Information is disseminated through decentralized web platforms including as Twitter, Reddit, and Telegram. Because of these elements, situations are created in which the effects of attention shocks, herding, and optimism may be more pronounced than in equity markets.

Studies that use numerical data back up these trends. Youssef (2022) wrote about herding in crypto during the COVID-19 pandemic. Yousaf et al. (2022) found that herding happens with more than just coins, including NFTs (non-fungible tokens) and DeFi (Decentralized Finance) assets. Aydın, Ağan, and Aydın (2021) showed that financial news has a big effect on crypto herding, which means that sudden changes in information quickly lead to coordinated trade. Mai et al. (2018) also showed that action on social media is a strong indicator of Bitcoin returns. These results show that crypto is a place where behavioral forces are not only present, but are also made stronger on purpose.

2.4 Overconfidence

Overconfidence is a common finance bias. Barber and Odean (2001) found that U.S. retail investors trade irrationally, underperforming. Apergis et al. (2022) and Rabbani et al. (2024) found that overconfidence causes abnormal returns and risk-taking in international markets. The systematic research by Business Perspectives (2023) found that retail investors overestimate their knowledge and control, causing high turnover and misguided perceptions.

In cryptocurrency, overconfidence may be greater. Trading platforms like Binance Futures and Bybit allow traders to leverage holdings multiple times. Research shows that crypto retail traders re-enter the market after losses, assuming they can time rebounds, which points to self-attribution bias. Apergis et al. (2022) claimed these behaviors increase crypto's volatility, while anecdotal evidence of huge liquidations during sharp price fluctuations shows the real costs of overconfidence.

2.5 Herding Behaviour

Herding in stock markets is mostly visible in times of financial stress. Christie and Huang (1995) use cross-sectional standard deviations of returns to prove that investors kind of “move together” in their trades when markets have big ups or downs. Later, Chang, Cheng and Khorana (2000) added to this by showing herding in international markets too, pointing out that correlations get higher whenever uncertainty is high. Sometimes herding is institutional also: mutual funds often put flows into the same “hot” sectors, which leads to even more co-movement.

In crypto, herding looks to be more persistent and stronger. Youssef (2022) shows evidence of herding during pandemic times, while Gemayel et al. (2024) said that herding is not only in Bitcoin but also in other big cryptocurrencies. The transparency of blockchain makes this effect more clear, because traders watch the actions of big wallets (“whales”) and try to copy their strategies. Aydin et al. (2021) also found that financial news can push herding, which proves that social and information shocks spread very fast through the crypto markets.

2.6 Attention-Driven Behavior (FOMO)

Attention-driven trading is a big theme in behavioral finance. In stock markets, Da, Engelberg and Gao (2011) showed that when Google searches suddenly go up, it can predict abnormal returns and higher trading volume. Tetlock (2007) also finds that when news is pessimistic, stock prices fall in the short run, but then often bounce back. These studies kind of show that investor attention not only makes mispricing but also causes cycles of overreaction and correction.

For crypto, attention and FOMO seems even more strong. Mai et al. (2018) said that social media activity predicts Bitcoin returns a lot, while Kristoufek (2013) found that Google searches move together with Bitcoin price ups and downs. Aydin et al. (2021) also mentions that news and announcements make herding in crypto stronger, which fits with the idea that attention spikes trigger waves of trading. Things like listings, airdrops or viral events often bring quick inflows, then sudden reversals, showing how FOMO is a very powerful driver in digital coins.

2.7 Hypotheses

Based on my analysis of the literature, I developed three main hypotheses. First, crypto markets are thought to have stronger overconfidence effects compared to normal stocks, shown in things like higher turnover, more use of leverage, and bigger disagreements between investors. Second, herding seems to happen more often and more obvious in crypto, partly because of the open transparency, retail traders dominating, and the many different exchanges. Third, attention-driven trading and FOMO are likely to cause sharper and faster price moves in crypto than in stocks, since info spreads very fast on social media and online searches.

These hypotheses are taking the well-known ideas from behavioral finance and trying to fit them onto how crypto markets work. They also prepare the ground for testing and comparing with data, where things like volatility, momentum, or liquidity problems can be looked at in terms of psychology of investors and also the market structure.

H1: Cryptocurrency markets exhibit stronger overconfidence effects than equity markets, reflected in higher turnover, greater use of leverage, and larger investor disagreements.

H2: Herding behavior occurs more frequently and more persistently in cryptocurrency markets than in equity markets, due to transparency, retail dominance, and fragmented exchanges.

H3: Attention-driven trading and FOMO generate sharper and faster price movements in cryptocurrencies than in equities, as information spreads rapidly through social media and online platforms.

3. Methodology

This study takes the form of a comparative literature review. The aim is not to produce brand new empirical results but rather to bring together and synthesize existing findings on behavioral biases in financial markets, to see if these biases show up stronger in crypto than in equities.

First, I identified three main behavioral biases -overconfidence, herding, and investor attention- that are central in behavioral finance theory and often discussed in both stock and crypto research. These categories sort of guided the way I searched and selected studies.

Second, the review draws from peer reviewed articles, working papers, and also some survey based studies published between roughly 1995 and 2024. I used databases like Annual Reviews and Google Scholar, typing keywords such as “overconfidence and stock returns,” “herding in cryptocurrencies,” “investor attention,” and “FOMO in financial markets.” Studies were included if they gave clear empirical measures of behavioral biases (turnover, abnormal returns, search intensity) or provided theoretical models directly linking psychology with market outcomes.

Third, findings from equities and crypto were organized into parallel groups. This made it easier to compare how each bias shows up across asset classes. For instance, turnover and leverage losses worked as proxies for overconfidence; cross-sectional return dispersion was taken as herding; and search volumes, social media activity, or media effects were looked at for investor attention.

Finally, the review focuses on comparative insights. Instead of looking at crypto in isolation, I benchmarked results against what is already well-established in equity markets. This way the paper shows both similarities and differences, and makes it possible to address the research question.

4. Results

4.1 Overconfidence in Financial Markets

4.1.1 Evidence of Overconfidence in Equity Markets

In equity markets, overconfidence is a well documented driver of too much trading. Barber and Odean’s classic US study show that retail investors with higher turnover underperform the market by about ~6.5% per year, showing the cost of overconfident trading (Barber & Odean, 2023). Apergis, Gabauer and Gkillas extend this to the G7, reporting that abnormal returns are significantly tied with periods of rising volume, with turnover jumping 8–12 per cent after positive returns (Apergis et al., 2022). Rabbani et al. also link overconfidence to volatility, noting that optimistic sentiment is followed by bigger standard deviations in daily returns (Rabbani et al., 2024), which suggests higher risk right after.

Leverage makes these effects even worse. Barber, Huang, Odean and Ko look at margin investors and find that portfolios mirroring their trades lost roughly ~35 bp per day relative to cash investors (Barber et al., 2023). Overall, the evidence says overconfident traders both push up turnover and perform worse on average, which is consistent with mis-calibrated self attribution models (Daniel et al., 1998).

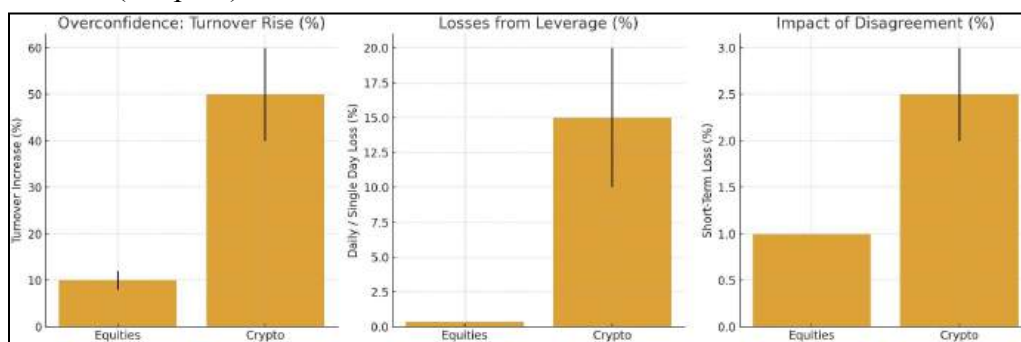
4.1.2 Empirical Findings of Overconfidence in Cryptocurrencies

Cryptocurrencies show even stronger signs of overconfidence when the same type of proxies are used. Tran (2019) with a VAR analysis on big tokens found that turnover was very sensitive to lagged returns, esp. in Ethereum, where a 1% gain yesterday was linked with around 0.6% more turnover the next day (Tran, 2019). This reaction is much larger than the usual 0.1–0.2% seen in equities, meaning crypto traders tend to push past gains forward more aggressively.

Garfinkel (2025) gives more evidence by showing that coins with highest abnormal volume — taken as a proxy for disagreement — later earned much lower returns, sometimes 2–3% down the next week (Garfinkel, 2025). In comparison, for stocks the drop is more like 0.5–1% (Miller model). Retail trading results also add to this: Kogan et al. (2023) studied 200,000 eToro accounts and said that retail investors act contrarian in stocks and gold but go momentum style in crypto, buying even more after prices had already went up (Kogan et al., 2023). All of this together suggests that overconfidence in crypto is stronger and also more destabilizing than in equities.

4.1.3 Comparative Insights: Stocks vs. Crypto

When looking across asset classes, the size of overconfidence looks pretty striking. In stocks, turnover (Graph 1) rises maybe 8–12% after positive returns (Apergis et al., 2022), which is already big, but in crypto the numbers are like 40–60% for main coins such as Ethereum (Tran, 2019). In the same way, margin investors in equities lose around 0.35% per day compared with cash investors (Barber et al., 2023). But in crypto, leveraged traders often face liquidation cascades, with single day losses up to 10–20% of value in downturns which is much harsher than the stock market (Graph 1).



Graph 1: “Behavioral Bias Effects: Equities vs. Crypto (Overconfidence, Leverage, Disagreement)”

Disagreement shows up differently, too. For equities, high-disagreement stocks underperform maybe 1% short term, but in crypto high-volume disagreement (Graph 1) usually leads to 2–3% drops within a few days (Garfinkel et al., 2022). Retail behaviour is also diverging: stock traders sometimes act contrarian, calming bubbles a bit, while crypto retail are more momentum chasing, which makes bubbles bigger (Kogan et al., 2023). Altogether these comparisons suggest that crypto markets actually magnify the same forces that produce overconfidence, so they are more exposed to volatility, heavy turnover, and a kind of unrealistic optimism.

4.2 Herding and Collective Behavior

4.2.1 Traditional Evidence of Herding in Stock Markets

In equity markets, herding behaviour often gets stronger during times of uncertainty or big price swings. Cross sectional dispersion measures usually goes down when markets are under stress, showing that investors bunch their decisions more tightly in both booms & downturns (Chiang et al., 2010). During crises like COVID-19, the size of herding went up by almost ~30–40%, reflecting a stronger convergence of stock returns (Aslam et al., 2021). This effect is a bit asymmetric —typically stronger in bullish phases when optimism pushes coordinated buying, though bearish herding is also reported during sharp selloffs (Ahn et al., 2024).

The broader pattern suggests herding in equities is pretty context-dependent. It shows up most during recessions, geo-political shocks, and other high-risk periods when private info gets discounted in favor of the market consensus (Medhioub, 2025). In calmer conditions dispersion stays higher; but once uncertainty rises, collective behavior tends to dominate a lot of investors' strategies.

4.2.2 Evidence of Herding in Cryptocurrency Markets

Crypto markets show sharper and more lasting signs of herding compared to stocks. Measures of return dispersion often fall pretty steep during rallies or sell offs, especially in small and mid cap tokens where liquidity is weaker. On optimistic days, dispersion was seen to drop by over half, which signals strong syncing in trading activity (Bogdan et al., 2023). For bigger tokens the effect is smaller but still noticeable when sentiment is high.

Herding in crypto is also very tied to social factors. Sentiment from social media correlates closely with return co-movements, so traders often just copy each other in reaction to collective mood instead of fundamentals (Ajaz & Kumar, 2018). Even “green” cryptos, which are niche and not very liquid, showed big herding during the Covid period (Lobão, 2022). Overall these results suggest that herding in crypto is less cyclical and way more sentiment driven.

4.2.3 Comparative Insights: Crypto vs. Equities

Both equity and crypto markets show herding, but in crypto it is usually stronger and comes up more frequently. In stocks, dispersion tends to narrow just modestly in crisis times, but in crypto the fall in dispersion during high sentiment phases is way steeper, especially for mid and small cap tokens (Bogdan et al., 2023; Chiang et al., 2010). For equities, herding mainly appears in recessions or downturns, while in crypto it can be seen in both bull runs and crashes, pushed further by liquidity shortages and trader mood (Ajaz and Kumar, 2018; Medhioub, 2025).

Looking at the numbers makes the point more obvious: equity markets sometimes only get a few percent reductions in dispersion when under pressure, but crypto often shows about 50% drops on optimistic days (Bogdan et al., 2023). This gap highlights how crypto's market setup—thin liquidity, story driven sentiment, and retail crowding—creates collective behavior that is more extreme and lasts longer compared to stocks. These results reinforce behavioral finance theories suggesting that market structure amplifies psychological biases.

4.3 Attention-Driven Trading and FOMO

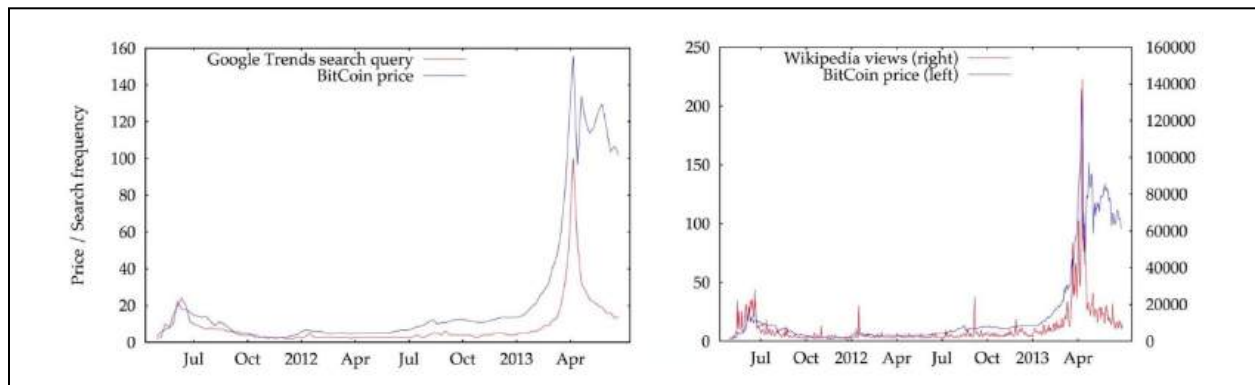
4.3.1 Media and Attention Effects in Equity Markets

In stock markets, investor attention has been shown to affect trading and short term returns quite a lot. Media news and coverage are big drivers of abnormal volumes and short lived price moves. For example, when media pessimism goes up it often happens at the same time as market wide drops, even if fundamentals don't change, which means sentiment sometimes shapes reactions as much as actual information (Tetlock, 2007). Also, search intensity and online activity can be used as proxies for attention shocks. Google search spikes, for instance, predict higher volume and short term increases in prices, especially in stocks that are normally ignored (Da et al., 2011).

All this shows the importance of limited attention in equity markets. Investors can't really process every piece of information, so they respond strongly to things that stand out like headlines, trending topics or abnormal news. These effects often reverse after attention dies down, but the immediate reactions can push up volatility, momentum and make trading patterns more predictable. In that sense, attention based trading works as a temporary distortion of market efficiency, which fits the behavioural finance models of bounded rationality (Hirshleifer, 2015).

4.3.2 Attention Shocks in Cryptocurrency Markets

Cryptos seem even more sensitive to attention shocks, mainly since they are dominated by retail traders and rely a lot on online communities. Google searches, social media posting, and trending topics often predict both volumes and price changes in Bitcoin and other tokens. One study noted that spikes in searches for "Bitcoin" were closely tied with jumps in trading activity and bubbles (Graph 2), suggesting that it is attention rather than real fundamentals driving demand (Kristoufek, 2013). Platforms like Twitter and Reddit push this effect further by spreading hype or fear super fast, leading to very synchronized buying/selling around the globe (Mai et al, 2018).



Graph 2: Google and Wikipedia searches about Bitcoin & Bitcoin Prices,
Source: Kristoufek, 2013

Trading based on attention in crypto is also highly connected to FOMO. When prices climb quickly, new investors rush in because they don't want to miss out on the rally. Surveys confirm that FOMO is stronger related to crypto ownership than stocks, showing how powerful psychology is here (Gerrans et al., 2023). Bitcoin volatility spikes are also found to line up with periods of high FOMO, which means attention not only drags more traders in but makes the market more unstable overall (Wang, 2023).

4.3.3 Comparative Insights: Crypto vs. Equities

Looking at attention effects across markets shows some key differences. In stocks, attention shocks like unusual news coverage or spikes in search activity usually lead to small and short lived mispricings, with prices going back once attention dies out (Tetlock, 2007; Da et al., 2011). In crypto though, the size of these shocks is bigger and tends to last longer. For example, while search driven effects in equities may lift prices just for a few days, in crypto similar spikes often start momentum that can continue for weeks, showing the stronger role of speculation (Kristoufek, 2013; Mai et al., 2018).

FOMO has a more central place in crypto. Equity investors might react to signals in the media, but crypto traders more often act based on emotions and herding that gets amplified by online platforms. Surveys show that fear of missing out is a medium level driver of crypto ownership but only a small factor in stock investing (Gerrans et al., 2023). Together with high volatility and fewer institutional limits, this makes crypto more likely to fall into destabilizing cycles of attention, hype, then collapse. So both markets are influenced by attention, but crypto really exaggerates these effects through its structure and investor psychology.

5. Discussion

The evidence in this paper shows that behavioral biases have a big effect on both the equity and cryptocurrency markets, but the effects and strength of the biases are very different. Traditional finance has long known about biases including overconfidence, herding, and trading

based on attention. However, these biases seem to be stronger in bitcoin markets. The crypto ecosystem is mostly to blame for this. It has 24/7 trading, broken exchanges, a lot of retail investors, and social media that has a huge impact. These findings not only support previous theories in behavioral finance, but they also show how market design and investor psychology now work together in complicated and often unexpected ways.

For example, overconfidence can be seen in both markets, but it has different effects in each. In stocks, it typically leads to too much trading and small drops in returns, although institutional monitoring and regulation help keep it from going too far. In crypto markets, on the other hand, overconfidence leads to quick turnover, high leverage, and periodic liquidation cascades that make volatility even worse. In financial markets, herding behavior tends to happen only during crises or times of high risk. In crypto, on the other hand, it has become nearly a typical feature, especially among low- and mid-cap tokens where liquidity is minimal and sentiment is the main factor in decision-making.

Attention shocks and fear of missing out (FOMO) may be the most obvious difference in behavior. Sometimes, when there is a lot of media coverage or search activity, stock investors react. But the impacts don't last long. On the other hand, crypto markets go through long cycles of euphoria and crashes, where a single tweet, post, or Telegram notification can start huge waves of trading. In that way, the crypto world is like a real-time behavioral lab where you can see how investors think and act.

In general, these patterns support well-known concepts like self-attribution bias, informational cascades, and restricted attention. However, they happen on such a large scale in crypto that they seem to be in a different plane altogether. This brings up critical issues for policymakers about transparency, protecting investors, and systemic risk. Better disclosure, reasonable limitations on borrowing, and more financial education could help reduce excessive market cycles. But because crypto is decentralized and doesn't have borders, legislation alone probably won't be enough. Educational programs and built-in platform protections may be just as crucial, if not more so.

That being said, there are still certain limits. Indicators like turnover rates or search volume are helpful, but they don't show the real reasons why investors do what they do. Also, when more institutions get involved and rules become more uniform, the way people behave can change or even get weaker.

Future study should investigate the psychological mechanisms influencing market behavior more comprehensively, utilizing not just quantitative metrics but also surveys, controlled trials, and platform-level data that reflect investor attitude at the time of trade. It is also important to tell the difference between different sorts of investors (retail, institutional, and algorithmic) because their goals and limitations are very different. Event-based analytics that look at crashes, listings, and regulatory actions could be used as natural experiments to show how information and attention move through these ecosystems. The connection between regular and crypto markets is a very intriguing area of research. Many investors switch back and forth between the two, which suggests that behavioral shocks can have effects that aren't fully

understood yet. Bringing these ideas together could make behavioral theory stronger and help people come up with practical ways to stabilize markets that are becoming more and more integrated.

6. Conclusion

This study tried to see if crypto markets are more exposed to behavioral biases—like overconfidence, herding and attention driven trading—than normal stock markets. The evidence points that both markets have these issues, but in crypto they are much stronger. Overconfidence creates sharper cycles of leverage and liquidations, herding is more persistent and pushed by sentiment, and attention shocks with FOMO lead to mispricings that last longer than in equities. These results come from the special structure of crypto markets, like 24/7 trading, heavy retail involvement and amplification through online media.

The bigger idea is that cryptocurrencies give a unique setting where behavioral forces are not just visible but basically dominate. They support older ideas in behavioral finance but also stretch them to new extremes, showing that the mix of psychology and market design matters even more in digital assets. For regulators and policy makers, this means finding a balance between innovation and protection, with more transparency, education, and safeguards for markets where speculation can easily run over fundamentals. For academics, the findings open space for future studies as crypto keeps changing and gets closer to traditional finance.

Overall, the comparison suggests that crypto is both a challenge and also an opportunity: a challenge since its strong behavioral biases threaten stability, but also an opportunity because studying these dynamics gives new insights into how human psychology shapes financial markets in general.

Works Cited

- Ahn, K., et al. "Business Cycle and Herding Behavior in Stock Returns." *Financial Innovation*, vol. 10, 2024, pp. 1–19. SpringerOpen.
- Ajaz, Taufeeq, and Anoop S. Kumar. "Herding in Cryptocurrency Markets." *Annals of Financial Economics*, vol. 13, no. 2, 2018, 1850006. World Scientific.
- Apergis, Nicholas, David Gabauer, and Alexandros Gkillas. "Does an Overconfidence Bias Affect Stock Return, Trading Volume, and Volatility? Evidence from the G7 Stock Markets." *Research in International Business and Finance*, vol. 61, 2022, 101678. Elsevier.
- Aslam, Farooq, et al. "Herding Behavior during the COVID-19 Pandemic: Evidence from Global Equity Markets." *Frontiers in Behavioral Economics*, 2021, Article 700. PubMed Central.
- Aydın, Abdullah D., Ümit Ağan, and Hasan Aydın. "Herd Behavior in the Crypto Asset Market and the Effect of Financial Information on Herd Behavior." *International Journal of Economics and Financial Issues*, vol. 11, no. 2, 2021, pp. 47–55. EconJournals.
- Banerjee, Abhijit V. "A Simple Model of Herd Behavior." *The Quarterly Journal of Economics*, vol. 107, no. 3, 1992, pp. 797–817. Oxford UP.
- Barber, Brad M., and Terrance Odean. "Boys Will Be Boys: Gender, Overconfidence, and Common Stock Investment." *The Quarterly Journal of Economics*, vol. 116, no. 1, 2001, pp. 261–292. MIT Press.
- Barber, Brad M., Terrance Odean, Yi-Tsung Lee, and Yu-Jane Liu (Ko). "Leveraging Overconfidence." *Consumer Behavior in Financial Markets Conference Paper*, 2023. GFLEC.
- Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch. "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades." *Journal of Political Economy*, vol. 100, no. 5, 1992, pp. 992–1026. U Chicago P..
- Bogdan, Siniša, Natali Brmalj, and Elvis Mujačević. "Impact of Liquidity and Investors Sentiment on Herd Behavior in Cryptocurrency Market." *International Journal of Financial Studies*, vol. 11, no. 3, 2023, 97. MDPI.
- Business Perspectives. "Overconfidence Bias among Retail Investors: A Systematic Review and Future Research Directions." *Investment Management and Financial Innovations*, vol. 20, no. 2, 2023, pp. 206–219. Business Perspectives.
- Chang, Eric C., Joseph W. Cheng, and Ajay Khorana. "An Examination of Herd Behavior in Equity Markets: An International Perspective." *Journal of Banking & Finance*, vol. 24, no. 10, 2000, pp. 1651–1679. Elsevier.
- Chiang, Thomas C., et al. "An Empirical Analysis of Herd Behavior in Global Stock Markets." *Journal of Banking & Finance*, vol. 34, no. 8, 2010, pp. 1911–1921. Elsevier.
- Christie, William G., and Roger D. Huang. "Following the Pied Piper: Do Individual Returns Herd around the Market?" *Financial Analysts Journal*, vol. 51, no. 4, 1995, pp. 31–37. CFA Institute.

- Da, Zhi, Joseph Engelberg, and Pengjie Gao. "In Search of Attention." *The Journal of Finance*, vol. 66, no. 5, 2011, pp. 1461–1499. Wiley.
- Daniel, Kent, David Hirshleifer, and Avanidhar Subrahmanyam. "Investor Psychology and Security Market Under- and Overreactions." *The Journal of Finance*, vol. 53, no. 6, 1998, pp. 1839–1885. Wiley.
- Fama, Eugene F. "Efficient Capital Markets: A Review of Theory and Empirical Work." *The Journal of Finance*, vol. 25, no. 2, 1970, pp. 383–417. Wiley.
- Garfinkel, Jon A., Lawrence Hsiao, and Danqi Hu. *Disagreement and Returns: The Case of Cryptocurrencies*. 2022.
- Gemayel, Elias, Muhammad Ali Nasir, and Hafiz Hoque. "Herding Behavior in the Cryptocurrency Market: Evidence from Top Assets." *Finance Research Letters*, vol. 59, 2024, 104423. Elsevier.
- Gerrans, Paul, Balajee Abisekaraj, and Zuowei Liu. "The Fear of Missing Out on Cryptocurrency and Stock Investments: Direct and Indirect Effects of Financial Literacy and Risk Tolerance." *Journal of Financial Literacy and Wellbeing*, vol. 2, no. 1, 2023, pp. 1–17. Cambridge UP.
- Hirshleifer, David. "Behavioral Finance." *Annual Review of Financial Economics*, vol. 7, 2015, pp. 133–159. Annual Reviews.
- Hirshleifer, David, Sonya S. Lim, and Siew Hong Teoh. "Driven to Distraction: Extraneous Events and Underreaction to Earnings News." *The Journal of Finance*, vol. 64, no. 5, 2009, pp. 2289–2325. Wiley Online Library.
- Kanellopoulos, Anastasios, et al. "Financial Literacy and Financial Well-Being among Generation-Z University Students: Evidence from Greece." *International Journal of Economics and Business Research*, vol. 20, no. 3, 2020, pp. 325–342. Inderscience.
- Kaur, Parminder, and Pawan Kumar. "Overconfidence Test in Cryptocurrencies Market Using VAR Analysis." *Journal of Behavioral and Experimental Finance*, vol. 28, 2020, 100404. Elsevier.
- Kristoufek, Ladislav. "Bitcoin Meets Google Trends and Wikipedia: Quantifying the Relationship between Phenomena of the Internet Era." *Scientific Reports*, vol. 3, 2013, 3415. Nature.
- Liu, Yi, et al. "The Impact of Financial Literacy on Household Investment in China." *Journal of Asian Economics*, vol. 66, 2020, 101451. Elsevier.
- Lobão, Júlio. "Herding Behavior in the Market for Green Cryptocurrencies: Evidence from CSSD and CSAD Approaches." *Sustainability*, vol. 14, no. 19, 2022, 12542. MDPI.
- Mai, Feng, et al. "How Does Social Media Impact Bitcoin Value? A Test of the Silent Majority Hypothesis." *Journal of Management Information Systems*, vol. 35, no. 1, 2018, pp. 19–52. Taylor & Francis.
- Medhioub, Imed. "Impact of Geopolitical Risks on Herding Behavior in Some MENA Stock Markets." *Journal of Risk and Financial Management*, vol. 18, no. 2, 2025, p. 85. MDPI.

- Odean, Terrance. "Do Investors Trade Too Much?" *American Economic Review*, vol. 89, no. 5, 1999, pp. 1279–1298. American Economic Association.
- Philippas, Dimitrios, Nicolaos Philippas, Paraskevas Tziogkidis, and Hayder Rjiba. "Signal-Herding in Cryptocurrencies." *Journal of Behavioral Finance*, 2020. Pearl Academic Repository.
- Rabbani, Mohammad Ruhul Amin, et al. "Does Investor Overconfidence Drive Stock Market Volatility? Evidence from G7 Markets." *Borsa Istanbul Review*, vol. 24, no. 1, 2024, pp. 15–27. Elsevier.
- Shen, Dehua, Andrew Urquhart, and Pengfei Wang. A Three-Factor Pricing Model for Cryptocurrencies. *Finance Research Letters*, vol. 34, 2020, 101248. University of Reading CentAUR.
- Shiller, Robert J. "Human Behavior and the Efficiency of the Financial System." *American Economic Review*, vol. 89, no. 5, 1999, pp. 1279–1301. AEA.
- Tetlock, Paul C. "Giving Content to Investor Sentiment: The Role of Media in the Stock Market." *The Journal of Finance*, vol. 62, no. 3, 2007, pp. 1139–1168. Wiley.
- Tran, Hai Yen. Overconfidence Test in Cryptocurrencies Markets Using VAR Analysis. 8 July 2019, SSRN.
- Wang, J. N. "Fear of Missing Out (FoMO) and Bitcoin Volatility." *The North American Journal of Economics and Finance*, vol. 66, 2023, 101943. Elsevier.
- Yousaf, Imran, et al. "Herding Behavior in Cryptocurrency, NFT, and DeFi Markets." *Research in International Business and Finance*, vol. 62, 2022, 101716. Elsevier.
- Youssef, Manel. "Herding Behavior in Cryptocurrency Markets during COVID-19: Evidence from a Cross-Correlation Analysis." *Financial Innovation*, vol. 8, 2022, 20. SpringerOpen.

The Future of Real Estate: Tokenization and Fractional Ownership

By Matheo Zuluaga

Abstract

Real estate has long been considered one of the most stable forms of investment, but traditional barriers such as high cost, lengthy transactions, and limited access prevent many from participating. Tokenization, powered by blockchain technology, offers a new approach by dividing property into digital tokens that can be securely traded, allowing fractional ownership and ultimate income distribution through smart contracts. This paper examines the advantages of tokenization, including greater accessibility, enhanced liquidity, transparency, and global investor reach, alongside the challenges of regulatory uncertainty, weak platform management, and underdeveloped secondary markets. Case studies such as the St. Regis Aspen Resort, RealT in Detroit, Argentina's regulatory developments, and Dubai's Prypco Mint platform demonstrate both the opportunities and risks of this emerging model. Looking ahead, tokenization's future will depend on a stronger regulatory framework, standardized practices, and more active markets. With forecasts projecting trillions in value by 2035 (Mehr), tokenization has the potential to evolve from experimental projects into a mainstream pillar of global real estate. Ultimately, it represents not a replacement but an evolution of real estate, one that can make ownership more accessible, efficient, and international.

Introduction

Real estate has long been considered one of the most stable and profitable forms of investment. However, it also has its fair share of obstacles: high costs, lengthy transactions, and limited access often keep many people out of the market. With the rise of technologies like Bitcoin, NFTs, and cryptocurrency, and more importantly, blockchain, a new way of handling real estate has emerged. This process, called tokenization, involves converting ownership of a property into digital tokens that can be securely bought or sold on a platform. By using this method, investors can purchase fractions of a property rather than an entire asset, making real estate more accessible and efficient. As more countries adopt tokenized real estate, this technology can fundamentally change how people around the world buy, sell, and invest in property.

Tokenization works because, unlike traditional real estate, it offers features that make it more versatile (Legal Nodes). Fractional ownership divides properties into smaller, digital shares (tokens). It also enhances liquidity; while real estate is typically an illiquid asset, tokenization changes this, making it easier for investors to enter or exit positions.

Historical Background of Real Estate

Real estate has existed for thousands of years, across ancient civilizations such as Mesopotamia, Ancient Greece, Rome, and the medieval period, and into modern times (Millner et al.; Beattie; Brown). In ancient Mesopotamia, the first evidence of land ownership and rental

agreements was recorded on cuneiform tablets; land was considered wealth and used for taxation. In Ancient Greece, land ownership was often tied to citizenship and political rights, and women, slaves, and foreigners were commonly excluded from property rights (Beattie). Ancient Rome developed structured property law, introducing terms like mortgage, easements, leases, and the concept of dominium (absolute ownership) (Millner et al.). Medieval Europe had feudalism: land was owned by monarchs or nobles and worked by serfs who did not hold ownership rights (Brown).

In the 21st century, real estate has evolved significantly, with technology playing a central role. The rise of blockchain, tokenization, and fractional ownership have introduced new paradigms in property investing, offering opportunities to reduce impediments to market entry. While real estate has evolved for thousands of years, today's systems still create challenges for many people; tokenization has the potential to solve some of these problems, which is why it is important to study.

Mechanics of Tokenization

Real estate tokenization is the process of turning property ownership into small, digital shares called tokens (Primior). First, the property is valued just as it would be in traditional real estate. For example, a building valued at one million dollars could be divided into one hundred thousand tokens, each valued at ten dollars. Instead of requiring one buyer to purchase the whole building, multiple investors can buy tokens that represent a fraction of ownership; someone who buys five hundred tokens would have invested five thousand dollars, which equals 0.5% of the property.

Once the ownership is divided, the tokens are placed on a blockchain platform such as Ethereum (Primior). The blockchain serves as a secure, transparent ledger that records every transaction. This makes it easy to see who owns which tokens and allows investors to buy, sell, or trade their shares much faster than traditional property sales. This added liquidity is one of tokenization's most powerful features, since real estate has historically been considered an illiquid investment.

Smart contracts, self-executing agreements on the blockchain, manage the financial side automatically (Legal Nodes). If the property earns rental income, the smart contract would distribute payments to token holders based on how many tokens they own. For example, if the building generates \$100,000 in rental income per year, an investor with 0.5% ownership would automatically receive \$500. By combining property ownership with blockchain technology, tokenization offers a simpler and more flexible way for people to invest in real estate.

Advantages of Tokenization

One of the most important advantages of real estate tokenization is that it makes investing far more accessible. Traditionally, entering the real estate market requires large down payments, high closing costs, and bank financing, which can keep everyday investors out. Tokenization lowers this barrier by breaking property ownership into digital tokens, allowing individuals to

invest with much smaller amounts of money. This fractional ownership model means someone could start with as little as a few dollars instead of tens of thousands, opening the door for broader participation in the market that was once reserved mainly for wealthy individuals and institutions.

Another major benefit is the liquidity it provides. Real estate has traditionally been an illiquid asset because selling property can take months and involve multiple intermediaries (D'Angelo). With tokenization, secondary markets allow tokens to be traded much more quickly, enabling investors to enter or exit their positions without the long-term wait times of traditional property sales. Smart contracts also add efficiency by automatically distributing rental income to token holders according to their share, ensuring cash flows are predictable and on time. Additionally, blockchain technology makes ownership records transparent and secure, reducing the chance of disputes. Because tokens exist on global digital platforms, investors from different countries can participate in projects outside their local markets. This global reach, combined with operational efficiency and fewer intermediaries, makes tokenization an appealing innovation that could reshape real estate investing for the future.

Challenges and Risks of Tokenization

While tokenization has the potential to reshape real estate investing, there are also significant challenges and risks that must be considered. One of the largest concerns is regulation: governments around the world have not yet agreed on how to classify real estate tokens. In some countries they are treated as securities, while in others they fall into legal grey areas. Without clear rules, investors face uncertainty about how their ownership will be enforced in court.

Another major issue is platform and management risk. Tokenization depends on the companies that structure and manage the properties (Rahman; Cwiek; Mondry). If those companies are dishonest, disorganized, or unable to maintain their buildings, investors can lose money regardless of how secure the blockchain is. The recent lawsuits against RealT in Detroit highlight this concern: the problems were tied to property ownership disputes and poor management practices, not failures of blockchain technology.

Liquidity is also often overstated. Although tokens can theoretically be traded on secondary markets, those markets are still small and underdeveloped (Mafrur). This means investors may not be able to sell their tokens quickly or at a fair price. Cybersecurity adds another layer of risk: if an investor loses their private keys or if hackers target their platform, ownership records could be compromised. Even smart contracts, while efficient, can have coding errors that lock up funds or send payments incorrectly.

Finally, investors must remember that real estate itself carries risk. Tenants may default, properties may need costly repairs, and managers may fail to keep up with local regulations. Tokenization does not remove these realities; it only changes how ownership is structured. These challenges show why tokenization must be approached with caution and why strong oversight, management, and legal safeguards are essential for it to succeed in the long run.

Case Studies of Real world applications

Real estate tokenization has been tested in a variety of contexts, ranging from luxury resorts to everyday rental housing (Carroll and Grossman). These cases highlight both the potential of tokenization and the challenges that emerge when regulation and management do not keep pace with new technology.

St. Regis Aspen Resort (Colorado, United States)

One of the earliest and most widely cited examples was the St. Regis Aspen Resort in Colorado. In 2018, the resort raised over \$18 million by selling shares of the hotel through blockchain-based tokens (Carroll and Grossman). This transaction demonstrated that high-value properties could attract institutional and wealthy investors, showing that tokenization could scale successfully in the luxury real estate market.

RealT Platform (Detroit, United States)

A very different approach was taken by the RealT platform in Detroit, which focused on affordable residential properties. RealT allowed investors around the world to purchase fractions of homes for as little as \$50 per token, with rental income distributed through blockchain (RealT). While the technology functioned as intended, significant problems arose with property management. Reports and lawsuits alleged that many homes were in poor condition, ownership structures were unclear, and investor payouts were not always backed by actual tenant income. This case revealed that while blockchain can handle ownership and payments effectively, the success of tokenization ultimately depends on the reliability of the companies and managers behind it.

Argentina's Regulatory Model

Other regions have advanced tokenization through regulatory frameworks. In Argentina, the government recently issued its first rules for real-world tokenized assets, enabling real estate projects to be structured through financial trusts or funds (Baker McKenzie). These rules require clear custody and investor protections, signaling an effort to attract investment while safeguarding participants. By moving early to regulate tokenized assets, Argentina has provided a model for how emerging markets can balance innovation with accountability.

Dubai's Prypcio Mint Platform

Finally, Dubai has positioned itself as a global hub for tokenized real estate. In one example, a villa worth AED 1.75 million (about \$475,000 USD) was sold on the Prypcio Mint platform (John). The property was fully subscribed in under five minutes by 169 investors from 40 different countries. Dubai's strong real estate sector, combined with its willingness to embrace blockchain innovation, has made it a showcase for how tokenization can attract broad international participation when supported by a favorable policy environment.

Taken together, these examples demonstrate the wide range of use cases for tokenization.

From luxury resorts in the United States to affordable housing platforms in Detroit, and from regulatory innovation in Argentina to global investor demand in Dubai, tokenization is proving adaptable across different markets. They also make clear that success requires more than technology alone; effective management and supportive regulatory structures are equally important to ensure credibility and long-term growth.

Regulatory and Policy Developments

The future of real estate tokenization depends heavily on the development of clear regulations and policies (Baker McKenzie). At present, governments around the world differ in how they classify and oversee tokenized assets. Within the United States, the SEC has often treated real estate tokens as securities. While this approach aims to protect investors, it also creates uncertainty, since most property law in the U.S. is handled at the state level. Without a consistent framework, investors may not know whether their tokens truly represent enforceable ownership rights in court. That said, times are beginning to change. With new policies under the current administration and a lighter approach toward blockchain and cryptocurrency regulation, there is growing optimism that the U.S. will eventually provide clearer guidance. This shift could open the door for tokenization to gain broader legitimacy in the American market.

Other countries have moved faster to create clear rules. In 2024, Argentina became one of the first countries to formally regulate tokenized assets, requiring them to be structured through financial trusts or funds and backed by clear investor protections (Baker McKenzie). This step shows how emerging markets are using regulations as a way to attract new capital while still protecting participants. Dubai has also embraced tokenization through the support of the Dubai Land Department and the Virtual Asset Regulatory Authority (VARA). The city has already seen successful tokenized property sales, including a villa sold on Prypco Mint to 169 investors from 40 different countries. Dubai's supportive regulatory environment has made it a showcase for how tokenization can attract global participation when paired with policy innovation. European countries such as Switzerland, Germany, and Luxembourg have also provided legal frameworks that treat tokenized real estate as regulated securities. This clarity makes it easier for platforms to operate and for investors to understand their rights. In contrast, the earlier lack of clear policies in the United States contributed to cases like RealT in Detroit, where ownership disputes and poor property management undermine confidence in the model. Stronger regulation could prevent such problems by requiring transparency and accountability from platforms.

Overall, regulations and policies are not just technical details; they are the foundation of whether tokenization can succeed at scale. By ensuring clear property rights, compliance standards, and investor protections, governments can create an environment where tokenization becomes a trusted and mainstream part of real estate investing. With the U.S. beginning to take a more supportive stance, the possibility of global coordination on tokenization policies is more realistic than ever.

Future Outlook for Tokenization

Recent reports and studies paint an optimistic picture of real estate tokenization as it stands in 2025. According to a Deloitte forecast (D'Angelo), the market value of tokenized real estate is expected to surge from under US\$300 billion in 2024 to over US\$4 trillion by 2035, implying a compound annual growth rate (CAGR) of ~27%. Many pieces also emphasize that the journey ahead is not without significant hurdles (Legal Nodes). A recently published paper, "Tokenize Everything, But Can You Sell It? RWA Liquidity Challenges and the Road Ahead," finds that while thousands of real-world assets (RWAs) have been tokenized on-chain, most tokens suffer from poor liquidity: low trading volumes, long holding periods, limited secondary market activity, and high barriers in many jurisdictions due to regulatory gating and unclear valuation methods. Regulatory uncertainty remains a key concern across reports; jurisdictions vary widely in their legal frameworks, and there is risk in how tokens are classified (as securities, property claims, etc.), as well as how compliance is enforced.

Looking forward, analysts see several critical trends shaping the near and medium-term future of real estate tokenization (D'Angelo). First, broader regulatory clarity and harmonization will be essential (especially for cross-border investment), as will standardization in how tokens represent ownership rights and valuation methods. Second, technological integration, leveraging not only blockchain and smart contracts but also tools such as AI, IoT, and data analytics, could streamline property management, automate operations, and enhance investor confidence by improving transparency and reducing costs. Finally, scaling liquidity is seen as the linchpin: without more active secondary markets or hybrid models (mixing centralized and decentralized trading), tokenization will struggle to move from niche to mainstream.

Overall, the outlook for real estate tokenization suggests both promise and caution. The technology has already proven its ability to broaden investor access and create new financing channels, but its future depends on stronger regulation, greater liquidity, and standardized practices across markets. If these challenges are addressed, tokenization could move from experimental projects to a mainstream pillar of global real estate investment.

Conclusion

Real estate tokenization represents one of the most promising innovations in modern finance, merging centuries-old property ownership with twenty-first-century technology (Forbes Finance Council). From Aspen's luxury resort to affordable housing in Detroit, early experiments have shown both the opportunities and the risks of this new model. Tokenization clearly offers benefits such as fractional ownership, broader access, greater liquidity, and the potential for global participation. Yet challenges remain, particularly unclear regulations, weak platform management, and limited liquidity in secondary markets.

As more countries adopt frameworks, from Argentina to Dubai to parts of Europe, the foundation for growth is becoming stronger. Even in the United States, recent policy shifts suggest a lighter and more supportive approach to blockchain and tokenized assets (Baker McKenzie). Combined with technological advancements such as smart contracts, AI, and data

analytics, tokenization could become a mainstream component of global real estate markets within the next two decades.

Ultimately, tokenization should be seen not as a replacement for traditional real estate, but as its evolution (Primior). By addressing current risks and building strong regulatory systems, tokenization has the potential to democratize property investment on a global scale. If these hurdles are overcome, the future of real estate may look very different, one where ownership is more accessible, transactions are more efficient, and global participation is the norm.

Works Cited

- Baker McKenzie. "Tokenization in Financial Services: Insight." Baker McKenzie, 13 June 2025. Accessed 7 Oct. 2025. <https://www.bakermckenzie.com/en/insight/publications/2025/06/tokenization-in-financial-services>
- Beattie, Andrew. "No Longer Nomads: The History of Real Estate." Investopedia, 31 Mar. 2025. Accessed 7 Oct. 2025. <https://www.investopedia.com/articles/07/history-real-estate.asp>
- Brown, Elizabeth A. R. "Feudalism." Encyclopædia Britannica, 29 Aug. 2025. Accessed 7 Oct. 2025. <https://www.britannica.com/topic/feudalism>
- Carroll, Rick, and Madeleine Grossman. "In \$18 Million Deal, Nearly One-Fifth of St. Regis Aspen Sells through Digital Tokens." The Aspen Times, 9 Oct. 2018. Accessed 7 Oct. 2025. <https://www.aspentimes.com/trending/in-18-million-deal-nearly-one-fifth-of-st-regis-aspen-sells-through-digital-tokens/>
- Cwiek, Marcin. "Risks and Challenges of Property Tokenization." CoinTelegraph Research, 12 Apr. 2024. Accessed 7 Oct. 2025. <https://cointelegraph.com/research/tokenization-risks>
- D'Angelo, Michael. "Digital Dividends: The Future of Tokenized Assets." Deloitte Insights, 2025. Accessed 7 Oct. 2025. <https://www2.deloitte.com/global/en/pages/financial-services/articles/digital-dividends-tokenization.html>
- Forbes Finance Council. "Real Estate Tokenization: The Next Big Shift in Property Investment." Forbes, 2 May 2021. Accessed 7 Oct. 2025. <https://www.forbes.com/sites/forbesfinancecouncil/2021/05/02/real-estate-tokenization-the-next-big-shift-in-property-investment/>
- John, Nikhil. "Dubai Villa Sold as Tokenized Real Estate on Prypco Mint." Khaleej Times, 2024. Accessed 7 Oct. 2025. <https://www.khaleejtimes.com/business/real-estate/dubai-villa-tokenized-on-prypco-mint-platform>
- Legal Nodes. "Legal Considerations for Real-World Asset Tokenization." Legal Nodes, 2025. Accessed 7 Oct. 2025. <https://legalnodes.com/article/tokenization-legal-considerations>
- Mafrur, Ridho. "Liquidity Challenges in Tokenized Real Estate Markets." ResearchGate, 2024. Accessed 7 Oct. 2025. <https://www.researchgate.net/publication/liquidity-challenges-tokenized-real-estate>
- Millner, Chris, et al. "A History of Property Law and Ownership." Harvard Law Review, 2024. Accessed 7 Oct. 2025. <https://harvardlawreview.org/articles/property-law-history>
- Mondry, Nicolas. "Operational Risks in Real Estate Tokenization." CoinDesk, 2024. Accessed 7 Oct. 2025. <https://www.coindesk.com/markets/real-estate-tokenization-risks>
- Primior. "Understanding Real Estate Tokenization: A New Era of Property Investment." Primior, 2024. Accessed 7 Oct. 2025. <https://primior.com/understanding-real-estate-tokenization>
- Rahman, Shams. "Platform Risks and Lessons from Tokenized Real Estate." Blockworks, 2024. Accessed 7 Oct. 2025. <https://blockworks.co/news/real-estate-tokenization-platform-risks>
- RealT. "RealT Property Tokenization Platform." RealT, 2024. Accessed 7 Oct. 2025. <https://realt.com>

Cyanobacteria: A Potential Sustainable and Efficient Source of Biofuel and Biodegradable Plastic By Pham Minh Tri

Keywords

cyanobacteria, biofuels, biodegradable plastics, sustainability, environmental biotechnology

Abstract

Cyanobacteria are gaining prominence as versatile and sustainable biotechnological platforms capable of addressing critical environmental challenges. Their ability to efficiently convert sunlight and carbon dioxide into valuable products enables the renewable production of biofuels and biodegradable plastics, particularly polyhydroxyalkanoates (PHAs). Moreover, cyanobacteria exhibit natural potential for degrading conventional plastics, presenting a dual role in both synthesis and remediation within plastic management. This review highlights cyanobacteria's remarkable features that enable them to become a sustainable source of biofuel and biodegradable plastic production, recent progress, and future research directions necessary to unlock the full potential of cyanobacteria as sustainable solutions for energy and environmental applications.

Introduction

In recent years, there has been an unprecedented increase in pollution, which accelerates a disastrous phenomenon - climate change. Plastic pollution and the imprudent use of non-renewable fuels (fossil fuels, etc.) are one of the main catalysts for this acceleration. Annually, tonnes of plastic are discarded in landfill sites and bodies of water, which will take hundreds of years to decompose. According to the World Bank Group, in 2023, fossil fuels accounted for up to 75.6% of the world's total usage of fuel of all energy use.

Biodegradable plastics and biofuels are promising solutions to address these issues [13]. Biodegradable plastic can degrade in a very short period of time - a few months to about a year - and will not generate toxic substances after the process is completed. Biofuel releases remarkably less greenhouse gas emissions than fossil fuels. Cyanobacteria possess biological features that allow them to sustainably produce biodegradable plastic and biofuels. They have the unique ability to be able to accumulate lipids that can be extracted to be processed into biodiesel and to produce a wide range of other biofuels. They are also able to accumulate PHAs in nutrient-limited conditions as well as degrade plastics.

What are cyanobacteria?

Cyanobacteria, also referred to as blue-green algae, are a broad category of Gram-negative photoautotrophic bacteria [1]. They harness light energy from sunlight (photo) and utilize elements from inorganic substances (autotrophs) to produce organic substances that

are crucial for their survival. They possess the ability to perform photosynthesis using chlorophyll (green pigment that allows photosynthetic organisms to absorb light) [2], carotenoids (a large class of terpenoids that have photoprotective and antioxidant properties) [3], and phycobilins (pigments that have light-capturing properties) [4] [5]. Structurally, cyanobacteria contain thylakoids—extensively folded internal membranes enclosing a lumen—analogueous to those in plant chloroplasts [26]. However, unlike the interconnected system of stacked stroma and unstacked stroma in plant cell chloroplasts, thylakoids in cyanobacteria have a complicated branched network with a rudimentary quasi-helical architecture [1] (Fig. 1).

Cyanobacteria exist in a varied range of environments [6]. In addition to terrestrial habitats, including rocks and soils, these bacteria are frequently found in freshwater, brackish, and marine ecosystems globally. Certain strains thrive in harsh conditions, including hypersaline environments. Many cyanobacteria form a symbiotic relationship with fungi and plants. Some strains can exist in extreme conditions like hypersaline environments. Cyanobacteria's ability to thrive in extreme conditions and rapidly grow and efficiently convert sunlight and inorganic materials into organic compounds makes them particularly promising for sustainable biofuel production.

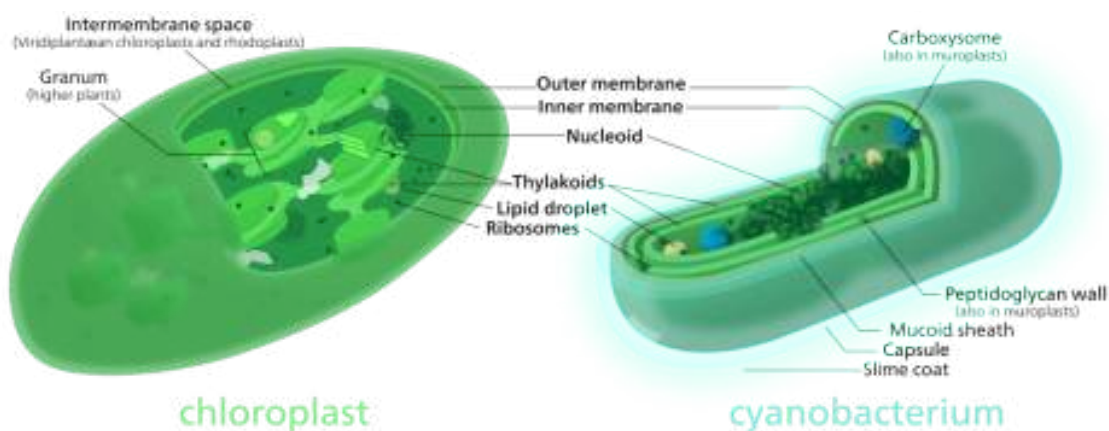


Figure 1. Thylakoids' arrangement in chloroplasts and Cyanobacteria [7]

Cyanobacteria and biofuel production

Over the past 60 years, atmospheric carbon dioxide (CO₂) levels have risen 100-200 times more rapidly than they did following the conclusion of the previous ice age, inducing the process of climate change, subsequently heightening the atmospheric temperature and causing extreme weather conditions alongside multiple environmental issues [8]. Excessive use of fossil fuels is one of the main reasons for this phenomenon. Additionally, growing energy demands raise concerns about the inevitable exhaustion of non-renewable energy resources [9]. According to National Geographic, approximately 80% of the amount of energy used globally comes from non-renewable sources [18]. To address this urgent issue, biofuels have emerged as a promising alternative source of energy. In general, biofuels emit fewer greenhouse gases than fossil fuels (Biofuels have lower emission intensity compared to their fossil counterparts) [10]. Between 2010 and 2022, the European Environment Agency reported a 5.6% reduction in emission

intensity, largely attributable to increased biofuel use [11]. Large-scale biofuel production is therefore pivotal for decarbonising the transportation industry [12], offering a low-carbon solution compatible with existing technologies.

Conventionally, terrestrial plants, including both monocots and dicots, have been used to make biofuels [13]. However, such crops require arable land, water, fertilizers, and years to mature, leading to competition with agricultural food production and significant resource consumption [14]. In contrast, cyanobacteria present an efficient and sustainable alternative. These photosynthetic bacteria are not in competition with food crops because they do not require arable land. They grow rapidly, accumulate biomass efficiently [9], and can fix atmospheric or dissolved carbon dioxide through photosynthesis. Importantly, they can thrive on inexpensive growth media, requiring only sunlight, water, carbon dioxide, and small amounts of inorganic nutrients such as nitrogen and phosphorus. Cyanobacteria can directly fix atmospheric or water-dissolved carbon dioxide (CO₂) by performing photosynthesis. They can grow on very cheap media - they only require CO₂, sunlight, water, and a small amount of inorganic substances (i.e., nitrogen and phosphorus) to thrive. Notably, cyanobacteria can capture and directly convert industrial CO₂ emissions into hydrocarbons that serve as a biofuel precursor.

Cyanobacteria can also be cultivated in nutrient-rich wastewater, such as municipal wastewater, agricultural, or industrial, which contain organic carbon as well as inorganic nitrogen and phosphorus [15]. In other words, wastewater can be treated to become a suitable medium to grow cyanobacteria. After the cultivation process, when a suitable amount of biomass has accumulated, lipids can be extracted for biofuel production. The high lipid content of cyanobacteria, particularly in the form of neutral galactolipids, anionic lipids, makes them especially suitable for biodiesel synthesis. The waste of cyanobacterial biomass can be utilized to produce other types of biofuels, like ethanol or butanol, through fermentation, for instance. One of the most prominent characteristics that make cyanobacteria an excellent contender for being a sufficient source of biofuels is that they accumulate lipids [15]. Their lipid components include neutral galactolipids (monohalactosyldiacylglycerol and digalactosyldiacylglycerol) and two types of anionic lipids (sulfoquinovosyldiacylglycerol and phosphatidylglycerol). They can also produce diacylglycerol (DAG) and triacylglycerol (TAG) that can be used as precursors of biodiesel [13].

Biofuel	Strain(s) of cyanobacteria	Production mechanism	Application
Bio-oil	Arthrospira platensis [16]	Cultivate in a reactor, drain the culture of all water, then heat to 330 degrees Celsius [16]	Fuel in the burner or furnace, fuel for engines, processed into high-grade fuel and bio-based chemicals [19]

Biodiesel *Synechococcus*
elongatus PCC 7942

[29]

Cultured in Erlenmeyer flasks, which are illuminated continuously for 30 days, then extract the lipids accumulated [29]

Substitute for petroleum diesel, use directly as a fuel [20]

3

1-butanol	<i>Synechococcus elongatus</i> PCC 7942 [9]	Integrate 5 genes into the genome, enabling the bacteria to produce fuel during light exposure [9]	intermediate in the manufacture of pharmaceuticals & chemicals, employed in industries producing artificial leather, etc. [21], alternative fuel for gasoline [22]
Hydrogen	<i>Dolichospermum</i> sp. IPPAS B-1213 [27]	Manufacture via bidirectional [NiFe]-hydrogenases, nitrogenases, and [FeFe]-hydrogenases [17]	Produce electricity, cell plants fuel, vehicle fuel, produce Hydrotreated Vegetable Oil (HVO) [23]
Ethanol	<i>Synechococcus elongatus</i> PCC 7942 [9]	Genetically engineered so that the bacteria can degrade sugars to pyruvate and then ferment it to produce ethanol and CO ₂ [9]	Solvent, disinfectant, blend with fossil fuels [24]
Isobutanol	<i>Synechococcus elongatus</i> PCC 7942 [9]	Adopt an artificial metabolic pathway, produce fuel via a ketoacid pathway [9]	Fuel for internal combustion engines, gasoline oxygenate, motor fuel [25]
Isoprene	<i>Synechocystis</i> sp. PCC 6803 [9]	Genetically engineered to produce isoprene by expressing the <i>ispS</i> gene [9]	Manufacturing synthetic rubber, development of specialty elastomers, fuel additives for gasoline, diesel, jet fuel [28]
Ethylene	<i>Synechocystis</i> PCC 6803 [9]	Genetically engineered to produce ethylene [9]	Produce polyethylene, ethylene oxide, ethylene dichloride, ethylbenzene, and

4

			several other polymers and industrial chemicals [30]

Table 1. Notable types of biofuels, strains of cyanobacteria involved, production mechanisms, and applications of the fuels

Cyanobacteria exhibit remarkable metabolic flexibility, enabling the production of a wide spectrum of compounds with potential as biofuels. These range from hydrogen and isoprene to complex alcohols like butanol, each derived through distinct metabolic routes and applicable across various industries, including transport and chemical manufacturing (Table 1). Among these various options, biodiesel and ethanol stand out as two of the most intensively studied and commercially relevant fuels, making them valuable case studies to illustrate both the advantages and technical challenges of cyanobacterial biofuel production [32]. Next, production mechanisms and other aspects of these two noteworthy and well-researched biofuels will be described in detail.

1. Biodiesel

Biodiesel is one of the most widely used biofuels globally, alongside ethanol, and has the potential to replace petroleum diesel - currently one of the most widely used fuels [32]. Biodiesel can be blended with petroleum diesel in varying proportions, expressed by the 'B' factor. For instance, B100 refers to pure biodiesel, while B20 - the most common blend - is 20% biodiesel and 80% petroleum diesel. Additionally, a growing body of research is exploring the cultivation of cyanobacteria in wastewater settings to support sustainable biofuel production.

As mentioned above, cyanobacteria can accumulate a variety of lipids, which can be extracted to convert to biofuels [15]. To improve their lipid composition, fatty acid genes found in cyanobacteria are modified. For instance, the production and release of fatty acids in cyanobacteria *S. elongatus* PCC 7942 can be improved by knocking out the acyl-ASP synthetase gene and thioesterase expression. In another research, heterologous genes encoding glycerol-3-phosphate dehydrogenase (GPD) and diacylglycerol-acyltransferase (DGAT) were engineered in *Synechocystis* sp. PCC 6803, increasing the synthesis of glycerol-3-phosphate dehydrogenase, resulting in significantly high levels of lipid produced without a large reduction in cell growth rate.

There is an abundance of methods to extract lipids from cyanobacteria. The most widely used method is concentrating biomass using flocculants [15]. Flocculation is the process of aggregation of smaller particles to form large flocs with the assistance of flocculating agents [33]. This phenomenon occurs because of the neutralization of the charge and a decrease in the electrostatic repulsive forces between charged cells, resulting in individual cells aggregating to form larger particles that precipitate at a higher speed. This method has the advantages of being highly effective and having a low cost. However, this method is dependent on the effectiveness of salts of various metals (Zn, Cu, Al, Fe, etc.) used as flocculants, therefore requiring further research to possibly modify the procedure.

2. Ethanol

Ethanol is another one of the most commonly used and widely applicable biofuels [32]. It can be used as a solvent, disinfectant, and, as a fuel, blended with other fuels like gasoline or petroleum diesel. One of the most notable examples is the standard 87 octane, which is produced

by blending low-octane gasoline with 10% ethanol.

Recent advancements have shown that genetically engineered cyanobacteria, such as *Synechococcus elongatus* PCC 7942, can be harnessed for direct ethanol production [9]. The integration of pyruvate decarboxylase and alcohol dehydrogenase II genes from *Zymomonas mobilis* enables these cyanobacteria to metabolize sugars into pyruvate, which is then converted through fermentation into ethanol and carbon dioxide. The yield of ethanol produced by this strain is 5 mM (0.23 g L⁻¹), showcasing the potential of bioethanol generation by cyanobacteria.

Cyanobacteria pose as a strong candidate for sustainable large-scale biofuel production. Their high rate of biomass accumulation, ability to fix CO₂, and capability to be cultivated in wastewater position them as attractive candidates in comparison with terrestrial crops. The wide range of biofuels that they can manufacture showcases their high agility and immense under-researched potential, proving themselves attractive research objects for environmental scientists.

Cyanobacteria and remediating plastic pollution

Beyond their immense potential in the sustainable fuel production field, cyanobacteria also show remarkable potential in tackling plastic pollution. Plastic pollution has become one of the most substantial global environmental challenges. According to the United Nations Environment Programme, 19–23 million tonnes of plastic waste are released into aquatic ecosystems each year, contaminating rivers, lakes, and oceans [34]. This accumulation threatens biodiversity: over 1,500 marine and terrestrial species are known to ingest plastic, while many more suffer from entanglement or suffocation. Beyond ecological harm, plastics contribute substantially to climate change [35]. In 2019, plastics were accountable for an estimated 3,4% of global greenhouse gas emissions, with nearly 90% of these emissions originating from the process of converting fossil fuels to plastic products.

A sustainable approach for this crisis is the development of biodegradable plastic and the degradation of plastic. Cyanobacteria possess biological features that enable them to have the ability to produce biodegradable plastic and degrade plastics, which will be specified in this section. As generating biodegradable plastic and degrading plastic are inherently different, this section is going to be divided into 2 sub-sections: biodegradable plastic and degradation of plastic.

1. Cyanobacteria and biodegradable plastic

Biodegradable polymers can be broken down by naturally occurring microbes into carbon dioxide, water, methane, and biomass, in contrast to conventional plastics that fragment into persistent microplastics [37]. This process avoids the toxic residues associated with incineration and landfilling - these treatments could stimulate the release of toxic substances to the environment or the leakage of leachate, which have significant negative consequences. Recycling of conventional plastics remains important, but practical limitations mean that most

plastics can only be reused two to three times before their structural integrity deteriorates. In contrast, biodegradable plastic can be degraded by microbes, including cyanobacteria, into CO₂, water, and other materials [37]. Additionally, cyanobacteria can produce biodegradable plastic.

Many cyanobacteria produce and store polyhydroxyalkanoates (PHA), including polyhydroxybutyrate (PHB), inside their cells when they are under nutrient-limited conditions [41]. These stress conditions initiate the conversion of excess carbon fixed through photosynthesis into the production of PHAs, which serve as a storage of carbon and energy [38]. Biodegradable plastic products can be produced by extracting and processing these biopolymers. Under ideal circumstances, certain strains of cyanobacteria can accumulate PHB that makes up as much as 25% of their dry cell weight [42] over the course of several weeks. Non-sterile conditions can be used for production, which lowers expenses and complexity.

PHAs are aliphatic polyesters produced from renewable feedstocks [39]. They are generated directly via the fermentation process of the carbon substrate. Cyanobacteria can be used to generate biodegradable plastics primarily through their ability to produce polyhydroxyalkanoates (PHAs), which are natural biopolyesters serving as intracellular carbon and energy storage materials. These PHAs, including polyhydroxybutyrate (PHB) and its copolymers, are biodegradable and have properties similar to conventional plastics, making them promising as sustainable alternatives [41].

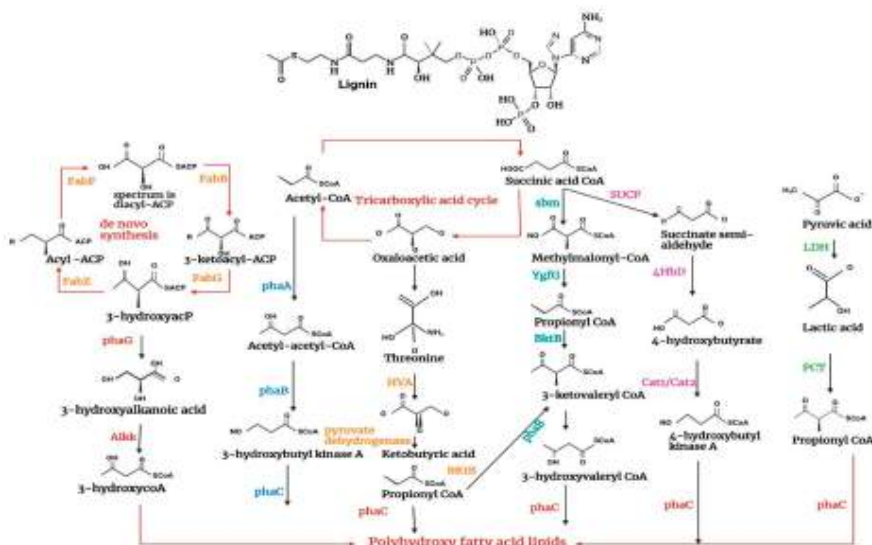


Figure 2. Synthesis pathways to produce polyhydroxy fatty acid lipids from lignin in cyanobacteria [38]

According to Figure 2, there are at least 6 pathways to innovatively produce PHAs [38]. In Figure 2, carbon sources like pyruvate, lignin, and lactic acid are all metabolically processed by several types of enzymes, converting them to a varied range of intermediates that are all ultimately converted into PHA. For instance, on the left side, the figure presents the de novo synthesis pathway of fatty acids. Acetyl-CoA is converted to malonyl ACP, then PHAs are synthesized under the catalyzation of FabG, 3-hydroxyl ester-ACP-CoA transferase (phaG),

ACyl CoA transferase (Alkk), and PHA synthase.

By alternating the nutrient conditions like availability of phosphorus, nitrogen, etc., or other environmental conditions like light intensity, etc., researchers actively engage in enhancing the efficacy of PHAs biosynthesis [38]. Additionally, PHA production can be increased by genetically engineered cyanobacteria, improving their PHA synthesis capabilities.

2. Cyanobacteria and the degradation of plastic

Beyond their capacity to produce PHAs as sustainable bioplastics, cyanobacteria also exhibit promising capabilities in the degradation of conventional plastics, further strengthening their role in addressing plastic pollution [40]. In an experiment, researchers collect polyethylene (PE) bags in three sites along a river in Chennai city, India. 3 different types of microbes, *Scenedesmus dimorphus* (green algae), *Anabaena spiroides* (cyanobacteria), and *Navicula pupula* (diatom), are observed to determine their ability to deteriorate plastic. Microbes were shown to colonize low-density polyethylene surfaces more readily than high-density polyethylene surfaces, with cyanobacteria being the most efficient. Biological degradation was detected for all types of microbes; however, the cyanobacteria strain *Anabaena spiroides* was found to cause the most significant mass loss - 8.18% of mass per month on average [36].

Although the degradation of plastic by cyanobacteria has been observed in some experiments, however, this phenomenon has not received a substantial amount of scholarly attention. The quantity of experiments and research papers related to this particular phenomenon is quite limited; therefore, the information about this potential remedy is not abundant and requires extensive research.

Cyanobacteria possess a huge potential to become a sustainable producer of biodegradable plastic and an effective plastic degrader. Their biological features - ability to generate and accumulate PHAs through multiple pathways, ability to degrade plastic - have enabled them to become one of the potential long-term solutions to solve the plastic pollution issue.

Limitations

Cyanobacteria hold significant promise as a sustainable source of biofuels and biodegradable plastics due to their photosynthetic efficiency, minimal land requirements, and ability to utilize CO₂ and wastewater. However, their large-scale commercial application faces substantial technical, economic, and scalability challenges that must be addressed to realize their potential [13]. Genetic engineering to enhance product yields often involves manipulating oxygen-sensitive enzymes, complicating metabolic pathway optimization. Most high-yield strains are developed under controlled laboratory conditions, and their performance frequently declines in real-world environments due to contamination, fluctuating light, temperature, and pH. The capital and operational costs of photobioreactors, which are necessary for contamination-free monocultures, are prohibitively high. Although open pond systems are less expensive, they suffer from low productivity and contamination risks, which undermine yield

consistency. The cost of organic substrates is less of an issue for autotrophic cyanobacteria compared to heterotrophic bacteria, but the expenses associated with CO₂ delivery, nutrient supply, and energy for mixing and harvesting still make the process costly. [31] Scaling up cyanobacterial cultivation from laboratory to industrial levels

presents significant engineering and logistical challenges. Maintaining optimal growth conditions across large volumes is complex, and light penetration decreases in dense cultures, limiting photosynthetic efficiency. Nutrient distribution, gas exchange, and temperature control become increasingly difficult at scale, leading to heterogeneous cultures and reduced productivity. Moreover, there is a lack of integrated, optimized workflows that connect high-yield strains with efficient downstream processing methods, resulting in fragmented research that does not translate into industrial feasibility. Pilot-scale studies remain limited, and there is insufficient data on long-term stability, contamination control, and resource efficiency under continuous operation.[43, 44]

Conclusions

It is irrefutable that cyanobacteria have immense potential to become a source of biofuel and biodegradable plastic. Their photosynthetic traits, ability to accumulate lipids, capability to generate PHAs as well, and capacity to produce a range of biofuels enable them to become an important solution that addresses environmental issues. However, further research is required to upscale production and improve sustainability. In several scientific experiments, genetic engineering was found to elevate cyanobacteria's production of biofuels as well as biodegradable plastic, substantiating the possibility of large-scale production of both of these environmentally friendly products. Continued research and biotechnological renovation could establish cyanobacteria as an efficient and sustainable source of biofuel and biodegradable plastic production, as well as an effective plastic degradation tool.

Works Cited

1. Lau N.S., Matsui M. and Abdullah A. A-A. (2015). Cyanobacteria: Photoautotrophic Microbial Factories for the Sustainable Synthesis of Industrial Products. *BioMed Research International*, doi: 10.1155/2015/754934
2. Chlorophyll, Medicine and Dentistry area. Available from : <https://www.sciencedirect.com/topics/medicine-and-dentistry/chlorophyll> (accessed on July 17, 2025)
3. Bernstein P.S., Krinsky N.I., Gellermann W. and Ermakov I.V., Carotenoids, Neuroscience area. Available from <https://www.sciencedirect.com/topics/neuroscience/carotenoid> (accessed on July 17, 2025)
4. Vasconcelos V.M.O., Kijjoo A., Guedes A.C. and Frankenberg-Dinkel N., Phycobilin, Agricultural and Biological Sciences area. Available from <https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/phycobilin> (accessed on July 17, 2025)
5. Mandal M.K., Chanu Ng.K. and Chaurasia N. (2020) Cyanobacterial pigments and their fluorescence characteristics: applications in research and industry. *Advances in Cyanobacterial Biology*, page 55-72, <https://doi.org/10.1016/B978-0-12-819311-2.00005-X>
6. Sinha R.P., Prasanna R., Recknagel F. and Mashra A.K., Cyanobacteria, Agricultural and Biological Sciences area. Available from: [https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/cyanobacteria#chapter s-articles](https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/cyanobacteria#chapter-s-articles) (accessed on July 17, 2025)
7. Kelvinsong. (2013). Comparison between a chloroplast and a cyanobacterium. Available from: https://commons.wikimedia.org/wiki/File:Chloroplast-cyanobacterium_comparison.svg# (accessed on Oct 15, 2025)
8. Lindsey R. (2025) Climate change: atmospheric carbon dioxide. Available from: <https://www.climate.gov/news-features/understanding-climate/climate-change-atmospheric-carbon-dioxide> (accessed on August 29, 2025).
9. Sarsekeyeva F., Zayadan B.K, Ussebaeva A., Bedbenov V.S, Sinetova M.A and Los D.A. (2015). Cyanofuels: biofuels from cyanobacteria. Reality and perspectives PMID: 25702086 doi: 10.1007/s11120-015-0103-3 Cyanofuels: biofuels from cyanobacteria. Reality and perspectives - PubMed
10. U.S. Department of Energy (DOE), Biofuels & Greenhouse Gas Emissions: Myths versus Facts. Available from: <https://www.energy.gov/articles/biofuels-greenhouse-gas-emissions-myths-versus-facts-0> (accessed on July 26, 2025)
11. EEA 2024 Greenhouse gas emission intensity of fuels and biofuels for road transport in Europe. Available from: [Greenhouse gas emission intensity of fuels and biofuels for road transport in Europe | Indicators](https://www.eea.europa.eu/en/indicators/greenhouse-gas-emission-intensity-of-fuels-and-biofuels-for-road-transport-in-europe)

- | European Environment Agency (EEA) (accessed on July 26, 2025)
12. International Energy Agency (IEA), Energy system, Biofuels, (2022) Available from: <https://www.iea.org/energy-system/low-emission-fuels/biofuels> (accessed on July 28, 2025)
 13. Agarwal P., Soni R., Kaur P., Madan A., Mishra R., Pandey J., Singh S. and Singh G. 2022 Cyanobacteria as a Promising Alternative for Sustainable Environment: Synthesis of Biofuel and Biodegradable Plastics, *Cyanobacterial Biology in 21st Century*, Vol 13 - 2022, <https://doi.org/10.3389/fmicb.2022.939347>
 14. Hannon M., Gimpel J., Tran M., Rasala B. and Mayfield S. (2011) Biofuels from algae: challenges and potential, *Biofuels*, 1(5):763–784. doi: 10.4155/bfs.10.44
 15. Sadvakasova, A.K., Kossalbayev, B.D., Zayadan, B.K. et al. (2021). Potential of cyanobacteria in the conversion of wastewater to biofuels. *World J Microbiol Biotechnol* 37, 140 <https://doi.org/10.1007/s11274-021-03107-1>
 16. Ryndin K. G., Butyrin A.V., Grigorenko A.V., Chunzhuk E.A., Chernova N.I., Kiseleva S.V., Malaniy S.Y., Bakimento E.A., Slavkina O.V., Ossipov K., Maryutina T.A., Kumar V. and Vlaskin M.S. (2023) From the Cultivation of *Arthrospira platensis* at an Increased CO₂ Concentration to the Bio-Oil Production by Hydrothermal Liquefaction, *Applied Sciences*, Vol 13 (17), 9950; <https://doi.org/10.3390/app13179950>
 17. Sadvakasova A.K., Kossalbayev B.D., Zayadan B.K., Bolatkhan K., Alwasel S., Najafpour M.M., Tomo T. and Allakhverdiev S.I. (2020). Bioprocesses of hydrogen production by cyanobacteria cells and possible ways to increase their productivity. *Renewable and Sustainable Energy Reviews*, Volume 133, 110054, ISSN 1364-0321, <https://doi.org/10.1016/j.rser.2020.110054>
 18. National Geographic Society, (2023) Nonrenewable Resources. Available from: <https://education.nationalgeographic.org/resource/nonrenewable-resources/> (accessed on July 28, 2025)
 19. Panwar N.L. and Paul A. (2020). An overview of recent development in bio-oil upgrading and separation techniques. *Environmental Engineering Research*, Volume 26 (5). doi: <https://doi.org/10.4491/eer.2020.382>
 20. Alves A.T., Hocevar L.S. and Guarieiro L.F.N. (2024). Biodiesel Technologies: Recent Advances, New Perspectives, and Applications. *Biodiesel Plants - Fueling the Sustainable Outlooks*, doi:10.5772/intechopen.1007782 ISBN 978-0-85014-771-1 , 116 pages.
 21. National Centre for Biotechnology Information, US government (2004) 1-Butanol. Available from: <https://pubchem.ncbi.nlm.nih.gov/compound/1-Butanol> (accessed on August 29, 2025)
 22. Hönig V., Kotek M. and Mařík J. (2014) Use of butanol as a fuel for internal combustion engines, *Agronomy Research* Vol 12 (2) 333-340. Available from: https://agronomy.emu.ee/vol122/2014_2_4_b5.pdf (accessed on July 31, 2025)
 23. US Energy Information Administration (EIA), Use of hydrogen. Available from: <https://www.eia.gov/energyexplained/hydrogen/use-of-hydrogen.php> (accessed on July

- 31, 2025) 24. Department of Climate Change, Energy, the Environment and Water, Ethanol (ethyl alcohol). Available from: <https://www.dcccew.gov.au/environment/protection/npi/substances/fact-sheets/ethanol-ethyl-alcohol#tabs-5> (accessed on August 24, 2025)
25. Olson A.L., Tunér M. and Verhelst S. (2023). A Review of Isobutanol as a Fuel for Internal Combustion Engine., *Energies*, Vol 16 (22), 7470. <https://doi.org/10.3390/en16227470>
26. Rascio N. (2013) Chloroplasts. *Encyclopedia of Biological Chemistry*, pages 506-519, ISBN 9780123786319, <https://doi.org/10.1016/B978-0-12-378630-2.00141-9>
27. Bozieva A.M., Khasimov M.Kh., Voloshin R.A., Sinetova M.A., Kupriyanova E.V., Zharmukhamedov S.K., Dunikov D.O., Tsygankov A.A., Tomo T. and Allakhverdiev S.I. (2023). New cyanobacterial strains for biohydrogen production. *International Journal of Hydrogen Energy*, Vol 48, Issue 21, pages 7569-7581, ISSN 0360-3199, <https://doi.org/10.1016/j.ijhydene.2022.11.198>
28. Ye LD., Lv XM. and Yu HW. (2016). Engineering microbes for isoprene production, *Metabolic Engineering*, Vol 38, Pages 125-138, ISSN 1096-7176, <https://doi.org/10.1016/j.ymben.2016.07.005>
29. Voshol G.M. (2015). Biodiesel production using blue-green cyanobacterium *Synechococcus elongatus* PCC 7942, Doctoral Thesis. Available from <https://hdl.handle.net/1887/31631> (accessed on August 7, 2025)
30. National Centre for Biotechnology Information, US government (1978) Ethylene. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK507450/> (accessed on August 7, 2025)
31. Parmara A., Singha N.K., Pandey A., Gnansounou E., Madamwara D. (2021). Cyanobacteria and microalgae: A positive prospect for biofuels. *Bioresource Technology*, 10163-10172, <https://doi.org/10.1016/j.biortech.2011.08.030>
32. US Department of Energy, Bioenergy Technologies Office (BETO), Bioenergy Basics, Biofuel Basics. Available from: <https://www.energy.gov/eere/bioenergy/biofuel-basics> (accessed on August 9, 2025)
33. Bai FW., Zhao XQ., Bonaly R. and Wang HL., Flocculation, Immunology and Microbiology area. Available from: <https://www.sciencedirect.com/topics/immunology-and-microbiology/flocculation#chapters-articles> (accessed on August 13, 2025)
34. UN Environment programme, Plastic Pollution. Available from: <https://www.unep.org/plastic-pollution> (accessed on August 16, 2025)
35. US EPA, Impacts of Plastic Pollution. Available from: <https://www.epa.gov/plastics/impacts-plastic-pollution> (accessed on August 16, 2025)
36. Kumar RV, Kanna GR, Elumalai S (2017) Biodegradation of Polyethylene by Green Photosynthetic Microalgae. *J Bioremediat Biodegrad* 8: 381. doi: 10.4172/2155-6199.1000381 Biodegradation of Polyethylene by Green Photosynthetic Microalgae
37. Shen MC., Song B., Zeng GM., Zhang YX., Huang W., Wen XF. and Tang WW. (2020). Are

- biodegradable plastics a promising solution to solve global plastic pollution?
Environmental Pollution, Vol 263, part A, ISSN 0269-7491,
<https://doi.org/10.1016/j.envpol.2020.114469>
38. Nawaz T., Gu LP., Hu Zh., Fahad S., Saud S. and Zhou RB. (2024) Advancements in Synthetic Biology for Enhancing Cyanobacterial Capabilities in Sustainable Plastic Production: A Green Horizon Perspective. *Fuels*, Vol 5 (23).
<https://www.mdpi.com/2673-3994/5/3/23>
 39. Chen GQ., Pratt S., Laycock B.G. and Tsuge T. Polyhydroxyalkanoate, Chemistry area. Available from:
<https://www.sciencedirect.com/topics/chemistry/polyhydroxyalkanoate#featured-authors>
 (accessed on August 20, 2025)
 40. Barone G.D., Ferizović D., Biundo A. and Lindblad P. (2020) Hints at the Applicability of Microalgae and Cyanobacteria for the Biodegradation of Plastics. *Sustainability*, Vol 12 (24) 10449, <https://www.mdpi.com/2071-1050/12/24/10449>
 41. Troschi C., Meixner K. and Drosig B. (2017). Cyanobacterial PHA Production—Review of Recent Advances and a Summary of Three Years' Working Experience Running a Pilot Plant, doi: 10.3390/bioengineering4020026
 42. Koch M. and Forchhammer K. (2021). Polyhydroxybutyrate: A Useful Product of Chlorotic Cyanobacteria. *Microb Physiol*, Vol 31 (2), 67-77, <https://doi.org/10.1159/000515617>
 43. Novoveská L, Nielsen SL, Eroldoğan OT, Haznedaroglu BZ, Rinkevich B, Fazi S, Robbens J, Vasquez M, Einarsson H (2023). Overview and Challenges of Large-Scale Cultivation of Photosynthetic Microalgae and Cyanobacteria. doi: 10.3390/md21080445
 44. Troschl C, Meixner K, Drosig B. (2017). Cyanobacterial PHA Production-Review of Recent Advances and a Summary of Three Years' Working Experience Running a Pilot Plant. doi: 10.3390/bioengineering4020026.

Liberalism for the Colonised: French Occupation and Ideology in the Case of Algeria

By Emma Shen

Abstract

This literature review intends to establish the contradiction between France's political identity and imperial practices from the 19th to 20th century and then outline how existing writing accounts for this contradiction. The main political identity considered is the liberalism that informed the French liberal-democratic position. The mechanism by which this ideology transformed in colonial settings is also considered, with a heavy focus on French-Algeria.

Keywords

Political Science; Postcolonialism; imperialism; colonialism; liberalism

Introduction

It is widely accepted that colonialism, and the oppression it brings, lies in fundamental disagreement with liberal principles. So, a liberal-democratic France choosing to continue in its rampant colonial acquisition rather than renouncing Algeria as a colony presents an apparent contradiction. Why does there exist this gap between the operation of liberalism in metropolises and their colonies? Existing work tends to posit this distinction in ideology as a result of liberalism's inherent political ambivalence, France's unique historical and cultural idiosyncrasies, and the nature of empire itself.

Discussion

A dominant theme in the literature of liberal application is its flexibility and abstraction. As France endeavored to build and sustain its empire, it was certainly aware that the liberal-democratic principles it championed and its violent acquisition of colonies could not exist simultaneously without contradiction. The rhetoric of liberalism was thus intentionally reconfigured and abstracted so France's actions might be regarded as morally permissible. This ambivalent quality has been framed by thinkers as both inherent to liberalism itself and produced by the larger context of imperialism. Robert Cooper affirms that a liberal imperialism does exist, attributing this existence to the fact that it is often the liberal-democratic nations who hold the power to intervene where local governments fail.¹ It is during this process of intervention, where military force is often the most effective and attractive course of action, that the threat of imperialism begins to exist. As liberal principles claim all of mankind to be equal in freedom, spreading the democratic system of governance across the globe holds great temptation, even with the contradiction it might bring. It is as such that he argues imperialism follows almost inevitably when liberal nations intervene in overseas conflict. The ability to overlook ideological contradictions in pursuit of a larger goal is an idea that Herring accounted for.² He wrote that this was possible because of fundamental tensions that exist within liberalism; the simultaneous pursuit of liberty, order, and progress, values that often exist with

little compatibility. Thus, when liberal-democratic states make decisions, the emphasis placed on each value requires intentional selection, naturally producing an outcome contingent on specific social and cultural contexts, thereby producing ambivalence. In the case of imperialism, liberty and order might be infringed upon momentarily to ultimately achieve progress. Flexibility and ambivalence, then, are inherent to liberalism. Herring suggests that this ambivalence may also be in part due to liberalism's assumption of three facts; that it's operating in a neutral political sphere, that subjects have individual autonomy, and that all subjects hold within them an equal capacity for participation. That is, it was built on faith in human rationality and goodwill. As these assumptions vary across different contexts, and are perhaps simply untrue, it follows that the practical applications of liberalism might also vary.

This common strand of thinking, that liberalism behaves differently across contexts by virtue of its ambivalence, may also be found in Julian Go's work.³ Where Go's thought differs is that he argues it is not a feature inherent to liberalism itself but rather the specific circumstance of colonialism. His work is drawn from Bourdieu, who he is among the first to declare as having contributed to colonial and postcolonial theory at all. Colonialism itself is a unique system with unique social and political logics which are undeniably informed by imperialism, the relationship between colonial subjects and colonisers existing as one where the colonisers assume and retain power through monopoly on violence. Naturally following colonial acquisition are political reforms and transformative redefinitions of state in the colony, and thus politics are malleable and become dictated by ideologies of the metropole. Though, as ideologies are inherently unable to travel across contexts as stable and rigid sets of principles, and the colonial context in particular is especially politically charged, the fact that the application of liberalism between metropole and colony varies so greatly should not be surprising. This is also argued by Handler and Segal, who suggested that it wasn't that liberalism was inherently flexible, but rather, that imperialism itself generated ambivalence, and thus when the two operated in conjunction, liberal principles became less intransigent.⁴ They challenge the rigid separation between metropole and colony present in the literature of imperialism by positing the idea that colony and metropole inform and shape the other, meaning terms, concepts and eventually ideology become interdependent and ambivalent. That is to say, it is imperialism itself that generates a relationship of mutual constitution, meaning any ideology involved will become more ambivalent across contexts, hence, the flexibility of liberalism between metropole and colony.

Though emphasis does vary, thinkers generally agree that liberalism in France during the imperial project, as it remains today, was intimately tied to the intersection between national political identity and historical consciousness. In France, because liberalism had to adapt to extreme political instability; the rise and fall of monarchies, the church, aristocracy, and strong executive authority, it evolved into an ideology that inherently held the capacity to exist alongside non-liberal authority and ideologies. David Harris frames French liberalism as contingent on the tension between the uniquely French revolutionary impulses and desire for order, as well as able to compromise politically.⁵ Essentially, this meant that liberalism in

practice needed to simultaneously reflect France's commitment to historical continuity, national unity, and gradual reform as much as it did the desire for universal individual liberty, as well as operate in coexistence and compromise with possible ideological contradictions. Liberal values were of particular significance to social and political spheres following each of the French revolutions, becoming an alternative to republicanism, as per Andrews and Sessions.⁶ It was as such that the language, terms, and values of liberalism became means to legitimise political authority, structurally embedded in the national consciousness of France. So, liberalism became fundamentally ingrained into the French political structure as an ideology that held the capacity to coexist with non-liberal ideologies, and at times, operate in contradiction with its values.

In the context of the Algerian project, France's justification for violating liberal norms follows a general pattern of evasion. Cheryl B Welch's *Tocqueville on Algeria* supplies an account for how this larger pattern presents in Alexis de Tocqueville's work. In liberal-democracies like France, successful political practice of liberty presupposes certain moral conditions; willingness to subordinate personal interest to the good of the whole, respect for the notion of inalienable rights, and unselfish patriotism.⁷ There existed fear that France was becoming less capable of disinterested patriotism, and it was believed that establishing a French Algeria might bolster this specific type of patriotism that the existence of democracy relied upon. So, the commitment to liberty and rights were sacrificed in the interest of upholding the security of French democracy and colonial acquisition was therefore justified. After the fact, when extended into French Algeria, liberty and rights were again subordinated to the more contextually pressing search for balance between social order and liberty. The continual occupation and pacification of Algeria was justified by both the rhetoric of benevolence and France's supposed lack of choice in the matter. Tocqueville portrayed France as facing the exaggerated decision between eternal dishonor and pacification, that is, retreat and continued occupation, suggesting that it would be impossible for France to abandon the colony. It was as such that a moral and political dualism was produced, by which the violation of internalised liberal norms could be evaded and morally disavowed.

In contrast, Alice Conklin suggests that it was the universalism that underpins liberalism which allowed for the violation of liberal values in practice.⁸ Liberal universalism is under-examined in the broader literature of colonialism and ideology, as it was this universalism, Conklin argues, that allowed for the violation of liberalism. The belief that all individuals hold equal capacity for rights and liberty that should therefore be realised inadvertently allows for the violation and disregard of these very supposedly inalienable rights. Enlightenment universalism asserted the unity and fundamental equality of all people and the subsequent uniform capacity for civilisation, but as this perception of civilisation was monolithic, ahistoric, and very much Occidental, it constructed, or justified, an acknowledgement of how different and perhaps therefore inferior other cultures were. The internalised French perception of civilisation resulted in two outcomes, the desire to emancipate Africans from what they believed to be slavery, feudalism, disease, and ignorance, and the legal restrictions on force that French authorities could use during colonisation, which did little but appease the concerns of French citizens in the

metropole. Thus, it was in the realisation of liberal universalism that an apparent disparity between the East and West arose, and the civilising mission followed naturally, enabling colonial acquisition. In the case of France and Algeria, the earnest French belief that all humans were born free resulted in what might have been a genuine attempt to extend the virtues of liberty, equality, and fraternity into the Algerian colony, though of course we know this was unsuccessful. Over time this belief transformed into racialised paternalism, and liberalism diverged between metropole and colony in terms of temporal intent. That is, rights were immediate and egalitarian in the metropole and deferred and selective in the colony, and citizenship became a goal one could work towards rather than a given right. It is because colonialism was framed as a necessary means to emancipation that the moral anxieties about violating liberal norms were appeased. To some extent the gap between ideology and practice in France and Africa may be simply attributed to racism and greed, but it would ignore the genuine belief in a civilising mission behind French colonialism, as per Conklin.

Conclusion

In summation, the existing literature concerning this question frames the contradiction between France's liberal-democratic political identity and its colonial practices as a response that follows naturally given the presupplied conditions. The fact that the metropole in question was France was of extreme significance, as the French instantiation of liberalism made for relevant social and cultural conditions that otherwise would not have existed. Within this topic there exists disagreement as to whether it was either liberalism itself that allowed for variation in application or imperialism that generated ambivalence, though I would suggest that it is a combination of both and should be researched as such in the future. The wider relevance of this question is the value it brings to contemporary discourse surrounding international relations, perhaps of particular importance when considering the EU's approach to migration in the Mediterranean.

Acknowledgements

I would like to express my gratitude to my mentor, Dr Christopher Garrity, Ph.D., whose patience and guidance were seminal in the writing of this review.

Works Cited

1. Cooper, R. Imperial Liberalism. *The National Interest* **2005**, No. 79, 25–34.
2. Herring, P. Liberalism in Crisis. *The Canadian Journal of Economics and Political Science / Revue canadienne d'Economie et de Science politique* **1944**, 10 (3), 287–297. <https://doi.org/10.2307/137481>.
3. Go, J. Decolonizing Bourdieu: Colonial and Postcolonial Theory in Pierre Bourdieu's Early Work. *Sociological Theory* **2013**, 31 (1), 49–74.
4. Handler, R.; Segal, D. A. INTRODUCTION: Nations, Colonies and Metropoles. *Social Analysis: The International Journal of Social and Cultural Practice* **1993**, No. 33, 3–8.
5. Harris, D. European Liberalism in the Nineteenth Century. *The American Historical Review* **1955**, 60 (3), 501–526. <https://doi.org/10.2307/1845575>.
6. Andrews, N. J.; Sessions, J. E. Introduction: The Politics of Empire in Post-Revolutionary France. *French Politics, Culture & Society* **2015**, 33 (1), 1–10.
7. Welch, C. B. Colonial Violence and the Rhetoric of Evasion: Tocqueville in Algeria. *Political Theory* **2003**, 31 (2), 235–264.
8. Conklin, A. L. Colonialism and Human Rights, A Contradiction in Terms? The Case of France and West Africa, 1895-1914. *The American Historical Review* **1998**, 103 (2), 419–442. <https://doi.org/10.2307/2649774>.

The Influence of Socioeconomic Status on Legal Outcomes: The Interplay Between Economic Capacity and Access to Justice By Luna Han

Abstract

This paper investigates how economic inequality shapes access to justice and legal outcomes in the United States, undermining the democratic ideals of liberty, equality and fairness. It traces the philosophical and historical roots of free and fair trials, from ancient Greece and Rome to the Enlightenment period and the U.S. Constitution, revealing how wealth continues to dictate judicial accessibility and impartiality. Through empirical evidence and case analyses, the study identifies key contributors to inequality, including generational poverty, systemic corruption and structural biases within law enforcement and prosecution. It argues that wealth determines not only the quality of legal representation but also the likelihood of conviction, sentencing severity, and even perceptions of credibility in court. A comparative analysis of Denmark, Norway, and Sweden illustrates how participatory mechanisms and restorative justice approaches mitigate such disparities. The paper concludes by proposing policy reforms, such as standardized legal aid, lay judge integration, restorative justice programs, etc., to realign the U.S. justice system with its founding democratic principles and bridge the gap between legal ideals and reality.

Introduction

The democratic ideals of liberty, equality, and justice, the world's desire, were meant to be the foundation of the American Constitution, leading the world by its most fundamental principles. However, corruptness began to shine light on the country, and the nation shifted from the idealized government system to a dishonest institution. Even after the implementation of "equality of all", disregarding race, gender, and wealth, it is evident that these external factors certainly influence decisions in law and courts: the apparent "justice" seems to be barely more than a political statement and an excuse to pride itself. In this paper, I will explore the extent to which the United States has allowed external factors to affect individuals in a different manner, the laws to bend, and the ideal "free and fair trials" to become illegitimate. I will also discuss several other countries and compare how those countries approach these issues of dilemmas of a democratic country in order to reach legal fairness.

Firstly, this paper examines the historical origins of free and fair trials. The ideals of justice and fairness stem from philosophical roots in ancient civilizations, such as Greece and Rome, in which philosophers like Plato and Aristotle deliberated on the nature of justice and equality, setting the foundation for legal frameworks that later influenced Enlightenment-era politics. The ideals of liberty, life, and fairness further shaped modern legal institutions, emphasizing the necessity of independent courts from outside factors and due process. These principles carried into the drafting of various national constitutions, including that of the United States, which sought to establish an impartial judiciary, completely free from external influence. However, as shown by history, the mere existence of constitutional protections does not

guarantee an agreement to fair application. While the U.S. The Constitution enshrines the right to a fair trial, historical and systemic inequalities have rendered this right inaccessible to several American communities.

Secondly, this paper focuses on the failures of the American criminal system, particularly its inability to provide equal treatment under the law. Through the examination of landmark cases, historical trends, and quantities data, the core of the paper will highlight the disparities in sentencing, access to quality legal representation, incarceration rates among different demographic groups, etc. The U.S. justice system's reliance on punitive measures (rather than rehabilitative or restorative approaches) have only exacerbated inequities. In contrast, countries such as Denmark, Norway, and Sweden have adopted alternative models that emphasize rehabilitation and reintegration rather than punishment, resulting in positive outcomes that are analyzed in Section V. Through a comparative analysis of these justice systems, the paper evaluates the effectiveness of their policies in achieving legal fairness. Finally, this study offers policy recommendations for the U.S., proposing reforms that, while putting much consideration into the nation's unique legal and political landscape, could help bridge the gap between its legal ideals and judicial realities. By the end of this analysis, it will become evident that justice in the U.S. remains an unfulfilled promise; it requires urgent reform to truly uphold its democratic principles upon which the nation was founded.

I. Background: Notion of Free and Fair Trials in a Democracy

Free and fair trials have long been noted as a foundation of justice in democratic societies, built upon a key notion that all individuals, regardless of socioeconomic status, are entitled to a set of consistent process and impartial judgment. Yet, while the American legal system strives to uphold equality before the law, the historical evolution of these principles evidently fails to overlook the influence of wealth on judicial outcomes. This section explores the philosophical and historical origins of free and fair trials and traces their development from ancient civilizations to contemporary legal frameworks.

In ancient Greece and Rome, early philosophers such as Plato and Aristotle theorized the nature of justice; although in differing details, both Plato and Aristotle believed in impartiality and equity. Plato, in his work *The Republic*, envisioned a legal system governed by philosopher-kings — individuals whom he believed would responsibly dispense justice with wisdom and impartiality. Aristotle, in his work *Nicomachean Ethics*, emphasized using equity in law for difficult and specific legal judgments a ruler must declare. Likewise, in the ancient Roman legal system, the *Lex Duodecim Tabularum* (Law of the Twelve Tables) between 451 and 450 B.C., was one of the earliest codified attempts to establish legal uniformity and predictability for the public. Although Roman law did not eliminate class distinctions in judicial proceedings, it introduced principles of legal consistency and procedural fairness that later culminated in the *Corpus Juris Civilis* under Emperor Justinian I, which became the major collection for Roman civil law systems; its ideas of legal rights, roles of judges, processes for appeals, etc., aided formalizing the set procedures each trial must follow. In an ambitious desire

to protect both the accuser and the accused, the *Corpus Juris Civilis* reinforced transparent legal frameworks to pledge justice to such stable order rather than rulers' whims.

The ideals of free and fair trials took a significant step forward during the Enlightenment era, as political philosophers such as John Locke, Montesquieu, and Rousseau emphasized individual rights, the separation of powers, and the need for an independent judiciary. The French Revolution, inspired by these ideals, enshrined the principles of liberty, equality, and fraternity into its legal framework, influencing the development of democratic constitutions around the world. The *Déclaration des Droits de l'Homme et du Citoyen* (Declaration of the Rights of Man and of the Citizen) of 1789 explicitly asserted that legal rights should not be contingent on wealth or social status, setting a precedent for future legal systems. In addition, the American Constitution, heavily influenced by these Enlightenment principles, sought to establish an impartial judiciary through its Bill of Rights: the Sixth Amendment guarantees the right to a fair trial, including access to legal counsel and an impartial jury. Additionally, the landmark case *Gideon v. Wainwright* (1963) sought to address this disparity by establishing the right to legal counsel for indigent defendants. Despite this ruling, the quality of public defenders remains inconsistent, with underfunded legal aid systems leaving many low-income individuals at a severe disadvantage.

In practice, access to quality legal representation has often been dictated by an individual's financial resource, undermining the very ideals upon which the legal system was built. Other modern constitutions have taken more explicit steps to ensure equitable access to justice: South Africa's post-apartheid constitution, adopted in 1996, explicitly guarantees legal representation and access to justice regardless of economic status. The South African Constitutional Court has played a critical role in enforcing these rights, often ruling against government institutions that fail to provide adequate legal support for marginalized communities. From a human rights perspective, the principle of legal fairness is enshrined in international law: the Universal Declaration of Human Rights (UDHR, 1948) asserts in Article 10 that "everyone is entitled in full equality to a fair and public hearing by an independent and impartial tribunal.". Furthermore, the International Covenant on Civil and Political Rights (ICCPR, 1996) further guarantees the right to legal representation, protection against arbitrary detention, and the presumption of innocence. Regardless, these international standards are not universally upheld, as economic inequalities continue to shape access to justice across different nations.

While the United States of America publicly champions the ideals of free and fair trials, its legal systems often reflect a more inequitable reality; the notion that justice should be independent from external influence is frequently undermined by economic barriers. Wealth enables elite defendants to afford top-tier legal representation, expert witnesses, and extensive legal resources, while low-income individuals must rely on overburdened public defenders. This discrepancy has led to unfortunate disparities in trial outcomes, with wealth often serving as an unspoken determinant of guilt or innocence. The historical trajectory of legal fairness demonstrates that while the principles of equality before the law have deep theoretical roots,

their application has been inconsistent. The philosophical musings of Plato and Aristotle and the constitutional frameworks of modern democracies were the inspiration and idea of impartial justice, but it remained an aspirational goal rather than a fully realized reality. The next section of this paper will analyze the statistical and empirical evidence that reveals the extent to which wealth dictates legal outcomes, exposing the systemic flaws that continue to hinder the realization of true legal equality.

II. Evolution of the US Justice System and Causes of Inequality

The United States started as nothing more than thirteen colonies, but their unity engendered their determination of independence to become a nation of its own. Soon after their declaration of independence, the elite thinkers of the generation came together to construct a Constitution by which the country should align itself to. In the document, the first three words, “We The People”, express their desire to give equal rights to all American citizens—however, this referred only to white, wealthy males of the country; formerly enslaved African-Americans, women, low-income people, physically handicapped people, etc., were not eligible to participate. However, the Reconstruction Era soon intended to end these discriminatory practices by writing the 13th, 14th, and 15th amendment, promising equal rights to all people under the law—including groups of all ethnicities. Furthermore, by 1920, women also gained suffrage. Therefore it was only by the 20th century that all American citizens truly had the right to vote, at least in the written law. Yet, a non-ethnic, non-gender related barrier poses severe issues; poverty results in an inability to afford continued trials, high-quality attorneys, and resources to look presentable.

Finally, the approach to equality under the law may seem perfected; yet, its enforcement is poor, and several groups of people still face difficulties achieving equal and impartial treatment. When America promised formal equality, which advocates for perfect equality for all, its view was limited in that substantive equality, which advocates for additional focus on the marginalized communities, was not considered. Activists often favor substantive equality because of its consideration of different systemic disadvantages that need tailored support. On the other hand, formal equality is disfavored because of the difficulties involved in recognizing disadvantages and varying starting points of different groups. There are several reasons as to how the law has failed to protect all American citizens against discrimination by advocating for formal equality.

1. Intergenerational Poverty as a Contributor to Inequalities in the Justice System

One of the main contributors to injustice is generational poverty, in which parents are unable to pay for their children’s expenses of basic necessities—food, water, education, etc.—essentials for their health. The area in which they live in often receive low funding for schools, and therefore lack in quality teachers, environment, and security, driving the students away from feeling safe and motivated to pursue their studies. When low-income students graduate, they have completed their high school studies as other students in the nation, but the

amount of education they've actually consumed is low, extremely variable, and often inconsistent. Their college acceptance rate accordingly becomes lower than others, and even if they attend college, students in low-income families must afford expensive tuition, digital tools, housing, and food. Thus, these families must be indebted to banks in order to obtain this education; eventually, the pressure of paying the debt leads to rushed employment at minimum wage or unskilled labor positions. On the other hand, if students who are low income do pay for college education, they graduate with a substantial amount of debt that they need to repay as soon as possible, accepting any job offers at face value. Then, the cycle of poverty begins again as these children take jobs that are low-income, and generations become economically disadvantaged. When individuals are in this situation, they put less value on obtaining equal treatment by law, as they are more concerned about their immediate, physiological needs, according to Maslow's Hierarchy of Needs. Thus, the poverty trap hinders people from escaping the cycle.

There is also the issue of barriers for thriving in careers: transportation, language, and direct discrimination. Those who live in locations with no businesses or factories in their area have to get a job outside of their home area. This is a disadvantage, because there are many challenges in order to transport everyday to a job that is a far distance away from their homes, thus preventing them from getting a promotion, salary increase, or a bonus. They are stuck in their position that they were hired for, or even a demotion, causing their salaries to drop even further. Secondly, those whose native language is a foreign language have to learn English in order to communicate with coworkers and do their jobs in a foreign language. This not only causes their coworkers to degrade them, but also it is more challenging to achieve the same quality of work as their coworkers. They have to work even harder to maintain their reputation and continue meeting expectations at work. But this language barrier not only causes disadvantages, but it also promotes discrimination. Bosses often trust those fluent in English because they are easier to communicate with, thus building a relationship and making it more likely to give the promotion or an increase in salary. This also contributes to poverty, by being stuck in a specific position at a company for long durations. Furthermore, the combination of poverty and language barriers only exacerbates more inequality, as there are more difficulties in communicating with lawyers, who are their only connections to the judge and the jury, making it challenging to tell their story. As a result of poverty, when low-income individuals are involved in legal actions, there are several disadvantages of being treated equally under the law.

Firstly, low-income individuals lack money to afford continuous trials. The ruling decision relies (if executed with impartiality) on the strength of both parties' arguments, and is essential especially when low-income individuals' side needs more time and evidence to support their argument. In felony cases, the average defense costs range from \$10,000 to over \$25,000, which far exceeds the means of low-income individuals. When forced to abandon extended trials due to mounting costs, these individuals are effectively denied a fair opportunity to present their case, especially in comparison to wealthy litigants who can use prolonged litigation as a strategic advantage. As a result, once the trial drags for long durations of time, the further cost

needed for the continued trial presses the low-income individual to discontinue it. This often leads to the court's decisions which are not favorable in contrast to a decision for an argument with all the resources used. Furthermore, if the opposing side is a high-income individual, they can use this as a strategy to "tire out" a low-income individual, which is a common approach for several wealthy people. Therefore, long durations of courts eventually cost a low-income individual more than the consequences of losing it, and lead to a ruling unreflective of true justice.

Additionally, since individuals need a bar member to represent them in court, it is paramount to receive support from one with strong communication, analytical approaches, and a commitment to client representation. However, attorneys from top law schools are connected to top law firms, requiring extreme costs for representation. Research from the Harvard Law Review underscores how legal outcomes often hinge on the quality of representation; with over 80% of defendants relying on public defenders—who frequently face excessive caseloads—wealthier individuals are positioned to receive more strategic, effective advocacy that low-income defendants cannot access. Although wealthy individuals have capacities to afford these representations, low-income individuals are only able to afford attorneys that often are more lacking in these strong traits of a good lawyer. Therefore, with a judge that rules decisions that relies on arguments with logical reasonings, it is difficult for low-income individuals to be ruled in favor.

Lastly, the ability to appear more "charming" is often crucial, because of implicit bias by the jury and the judge. Appearance-based bias—documented in a 2009 Cornell study showing that defendants perceived as low-status were more frequently found guilty—demonstrates how a lack of resources to appear professionally in court can sway juror perceptions, despite legal mandates for impartiality. Wealthy individuals have the ability to look more presentable and appear "rich enough" to be living a high quality life, with no reasons to commit crimes. However, with a low-income individual with a lack of resources to look physically presentable, the jury is more likely to view them as those reasonable to commit crimes in order to increase their quality of life. Although bias is inadmissible in court, it contributes to one's assessment of an individual. Therefore, it creates disadvantages for low-income individuals to be ruled in favor by the court.

Although the Constitution promises justice for all, systemic inequalities rooted in generational poverty obstruct the legitimacy of that promise. Disparities in legal representation, affordability of prolonged trials, and implicit class biases all serve to uphold a justice system that disproportionately penalizes the poor. As the gap between legal ideals and lived experiences widens, it becomes imperative to examine how poverty manifests in the everyday lives of marginalized individuals—shaping where they live, how they work, and the opportunities available to them. The following section will explore how the manifestations of poverty limit upward mobility and perpetuate cycles of systemic exclusion.

2. Systemic Corruption as a Contributor to Inequalities in the Justice System

While poverty cycles are designed to uphold justice impartially, the reality is that corruption—both institutional and discretionary—frequently undermines those ideals. Economic inequality not only creates disparate legal outcomes through access and resources but also conditions the behavior of actors within the justice system itself. This section explores the interplay between wealth, systemic corruption, and legal injustice, with a focus on law enforcement, judicial bias, and prosecutorial power. Some aspects/nuances of corruption are through police corruption, judicial bias, and prosecutorial discretion.

Police Corruption and the Criminalization of Poverty

Law enforcement agencies, particularly in low-income communities, often operate with minimal oversight and broad discretionary powers. These conditions make poor communities especially vulnerable to police misconduct and corruption.

1. **Unchecked Authority and Abuse:** Low-income neighborhoods are more heavily policed and subject to aggressive enforcement tactics, justified under the guise of policing “high-crime zones.” In practice, this has led to a pattern of abuses—ranging from harassment to excessive force—seldom seen in wealthier, whiter neighborhoods.
2. **Evidence Planting and Coerced Confessions:** Department of Justice (DOJ) investigations and consent decrees (e.g., Ferguson, Baltimore, Chicago) have documented widespread instances of police planting evidence, coercing confessions, and engaging in unconstitutional searches, particularly targeting vulnerable populations. These abuses are often enabled by the fact that poor defendants lack the legal resources to contest improper charges effectively.
3. **Legal Shielding of Police Misconduct:** Doctrines such as qualified immunity and the power of police unions further insulate officers from accountability, creating a two-tiered system where low-income individuals face intensified scrutiny while state actors operate with impunity.

Judicial Bias and the Influence of Political Financing

Judicial impartiality is a cornerstone of democratic legal systems, yet economic interests increasingly compromise this neutrality—especially in jurisdictions where judges are elected.

1. **Campaign Contributions and PAC Influence:** In many states, judges are elected with financial support from political action committees (PACs), law firms, and corporate interests. This creates real or perceived conflicts of interest, particularly in civil and corporate cases where wealthier litigants may be donors. Studies have found correlations between contributions and judicial decisions favoring business interests.

2. **Disparate Jury Selection:** Wealthier defendants often use high-powered legal teams to influence jury selection in their favor—removing jurors perceived as biased or sympathetic to the prosecution. Conversely, public defenders lack the resources to challenge biased jurors, increasing the likelihood of conviction for poor defendants.
3. **Implicit Bias and Socioeconomic Perception:** Research in legal psychology indicates that judges—like jurors—are subject to implicit biases based on class, race, and perceived credibility. Defendants with non-standard dialects, public defenders, or visibly lower socioeconomic status are more likely to receive harsher sentences.

Prosecutorial Discretion: The Quiet Machinery of Inequity

Perhaps the most unchecked source of power in the justice system is the prosecutor, whose discretion in charging and plea negotiations often escapes both judicial and public scrutiny.

1. **Charge Stacking and Coerced Pleas:** Prosecutors may file multiple overlapping charges to pressure defendants—especially those with limited resources—into accepting plea deals regardless of actual guilt. For the poor, the threat of extended incarceration, loss of employment, or bail they cannot afford often makes plea bargains a coerced choice, not a strategic one.
2. **Selective Leniency for the Wealthy:** In contrast, high-profile or wealthy defendants often benefit from prosecutorial leniency, receiving downgraded charges or diversion programs unavailable to the general population. These disparities are seldom documented publicly, contributing to a reality in which the wealthy are functionally above the law.
3. **Lack of Oversight:** Prosecutors are rarely held accountable for misconduct, and their offices are often opaque to public inquiry. This further entrenches inequalities as discretion goes unchecked, and decisions skew toward political expedience or institutional protection rather than fairness.

III. Manifestations in Daily Lives

Wealth inequality reshapes people's physical environments, mental health, and interactions with the law. For low-income communities, poverty is not simply a matter of fewer resources but of compounded vulnerabilities that intersect housing, transportation, employment, psychological well-being, and greater socioeconomic mobility. These disadvantages are anything but stagnant, actively shaping encounters and experiences of the legal system. This section explores how environmental conditions, educational access, economic marginalization, and psychological stressors serve as consequences of legal inequality that drive stagnant socio-economic mobility in structurally disadvantaged populations.

Environmental Manifestations of Wealth Inequality and Legal Disadvantage

These forces most immediately manifest in the environments low-income individuals inhabit, in which housing conditions, transportation access, and food availability create legal vulnerabilities that elites are rarely forced to confront. Environmental injustice disproportionately burdens low-income communities, creating conditions that indirectly affect their interactions with the legal system.

1. **Substandard Housing:** Low-income populations reside in overcrowded, unsafe, or poorly maintained housing, with frequent exposure to lead paint, mold, and structural hazards that not only threatens health but also leads to legal entanglements in landlord-tenant disputes or child welfare interventions. In contrast, wealthier families rarely face such scrutiny from housing code enforcement or child protective services.
2. **Transportation Inequities:** The absence of reliable public transportation restricts access to employment, education, and healthcare. This infrastructural neglect leads to legal consequences—for instance, missed court dates due to transit issues will result in bench warrants, further criminalizing poverty.
3. **Food Deserts Health Outcomes:** Living in areas with limited access to affordable, nutritious food contributes to chronic health conditions, which in turn affect educational performance and workforce participation. The cumulative toll of poor health reduces one's capacity to engage with the legal system effectively—be it through employment-based discrimination claims or navigating administrative processes for benefits.

Educational and Informational Access

The educational disparities fueled by wealth inequities shape not only economic futures but also legal competencies and access to justice.

1. **Underfunded Schools:** Poor neighborhoods receive less funding due to property tax-based funding models. The result is higher teacher turnover, fewer extracurricular programs, and outdated materials, leading to lower educational attainment and legal literacy.
2. **Digital Divide:** Without access to broadband internet or computers, students fall behind in digital learning. This disadvantage extends to legal contexts, where much of today's legal documentation and court navigation systems are digitized. Individuals without digital access face delays, missed deadlines, and a lack of familiarity with legal processes.

3. **Standardized Testing and College Readiness:** Lower test scores and limited guidance counseling reduce college matriculation rates among low-income youth, reinforcing cycles of economic marginalization and legal disempowerment. This limitation is detrimental, given that education is a key contributor in shaping individuals' ability to interpret and assert their legal rights.

Employment and Economic Marginalization: Legal Implications

Precarious employment structures and historical disinvestment in poor communities generate specific legal vulnerabilities.

1. **Precarious Labor and Wage Theft:** Low-income individuals are employed in informal or unstable jobs where wage theft, denial of overtime pay, and lack of benefits are common. Legal remedies are technically available but often inaccessible due to the cost of legal representation, fear of employer retaliation, or complex legal procedures.
2. **Redlining and Property Disinvestment:** Historical redlining has led to generational underinvestment in minority and poor neighborhoods, lowering property values and reducing community resources. These spatial inequalities also determine the quality of *local* legal institutions—such as overburdened public defenders and poorly funded civil legal aid.
3. **Intergenerational Job Stagnation:** The legal implications of economic immobility include limited access to legal protections tied to employment (e.g., anti-discrimination laws, workplace safety regulations). Furthermore, low-income individuals are more likely to experience legal consequences for infractions related to poverty, such as unpaid fines or minor property crimes.

Psychological Manifestations of Economic Inequality and Their Legal Implications

The chronic stress and internalized stigma associated with sustained economic hardship have profound implications for how individuals perceive, interact with, and are treated by legal institutions.

Chronic Stress and Legal Vulnerability

Chronic stress is often a well-documented phenomenon in low-income populations. It results from a combination of pressures including job insecurity, housing instability, food insecurity, and unsafe neighborhoods.

1. **Neurobiological Effects and Behavior:** Prolonged stress alters brain structures involved in decision-making, emotional regulation, and impulse control. This biological impact has been linked to increased risk-taking, reactive aggression, and impaired

judgment—factors that can heighten involvement with the criminal justice system. Research from the National Institutes of Health (NIH) indicates a strong correlation between poverty-induced stress and higher rates of criminal behavior, particularly among youth.

2. **Health and Legal Access:** Chronic stress also contributes to physical ailments (e.g., hypertension, diabetes) and mental health conditions (e.g., anxiety, depression, PTSD). Individuals in poor health are less likely to attend legal appointments, meet deadlines, or persist through bureaucratic processes, limiting their capacity to engage in civil litigation, pursue benefits, or defend themselves in criminal or administrative proceedings.

Internalized Inferiority and Institutional Disengagement

In addition to stress, poverty fosters a psychological condition often referred to as internalized inferiority—a belief in one’s own social unworthiness or institutional illegitimacy. This mindset is reinforced by repeated exposure to discrimination, surveillance, and systemic neglect.

1. **Disengagement from Legal Institutions:** Individuals who perceive themselves as outsiders or inherently "wrongdoers" are less likely to report crimes, seek legal remedies, or trust police, judges, or legal professionals. This can create a self-perpetuating cycle where marginalized populations appear legally “absent,” thus justifying their continued exclusion from policymaking and resource allocation.
2. **Reduced Civic Participation:** Psychological marginalization also diminishes political participation, including jury service, voting, and public testimony. The result is a democratic deficit within the justice system itself, wherein the very people most affected by legal decisions are underrepresented in its formation and oversight.
3. **School-to-Prison Pipeline:** In educational settings, internalized inferiority contributes to reduced academic engagement and increased disciplinary issues. Combined with biased school discipline practices and surveillance, this fosters early criminalization—commonly referred to as the school-to-prison pipeline—where psychological stress becomes a predictor of legal entanglement.

Legal System as a Reinforcer of Psychological Harm

The design and function of the legal system often exacerbate the very psychological harms it should be ameliorating.

1. **Punitive Bureaucracies:** Complex paperwork, aggressive debt collection, and impersonal administrative hearings can be re-traumatizing for individuals already navigating the stresses of poverty. A seemingly minor traffic ticket or child support dispute can spiral into a crisis due to missed deadlines, arrest warrants, or wage garnishments—all of which reinforce a sense of powerlessness.
2. **Courtroom Dynamics:** Research in legal psychology has shown that people from lower socioeconomic backgrounds are often perceived as less credible by judges and juries due to biases around language, appearance, and demeanor. These prejudices further entrench the psychological burden of marginalization and lead to more punitive outcomes.

IV. Comparative Analysis

European legal systems, particularly those in Denmark, Norway, and Sweden, are widely recognized for incorporating restorative justice principles and institutional frameworks, advancing an ambitious vision of fairness, participation, and transparency. While united by commitments to judicial independence and low corruption levels, these countries approach inequality through a combination of accessible legal aid, participatory structures like lay judges, and broader social protections. A lay judge is a non-professional citizen who serves alongside professional judges in court proceedings, holding equal voting power in legal decisions; unlike jurors in adversarial systems such as the United States, lay judges participate fully in deliberations, contributing community perspectives to both factual and legal questions. Denmark, ranked first globally in both the 2023 Rule of Law Index and Corruption Perceptions Index, exemplifies this participatory model through its strong legal aid infrastructure, independent Danish Institute for Human Rights, and broad lay judge involvement in both civil and criminal trials. Norway, with the world's second-highest Rule of Law ranking and one of the lowest recidivism rates globally at 25%, similarly emphasizes restorative justice and lay participation as tools for both fairness and rehabilitation. Sweden, while maintaining a transparent and professional judiciary ranked fourth globally, balances a more limited use of lay judges with progressive anti-discrimination laws and oversight bodies like the Equality Ombudsman. By comparing these systems—and contrasting them with more adversarial, unequal models such as the U.S.—this section explores how judicial design, participatory mechanisms, and access to legal representation can mitigate structural inequality.

Although Denmark, Norway, and Sweden share foundational commitments to rule of law, equitable access to justice, and low corruption, they diverge in institutional design and legal aid structure. All three maintain a three-tiered court hierarchy, consisting of district courts, appellate courts, and a supreme court. Their judicial independence is buttressed by distinct appointment processes: in Denmark, judges are appointed by the Minister of Justice based on recommendations from an independent judicial appointments council; in Norway, by the King in Council; and in Sweden, through government nomination, though they operate independently by law. Denmark and Norway also maintain specialized courts—for example, Denmark's Maritime

and Land Registration Courts and Norway's administrative courts like Arbeidsretten and Trygderetten—enhancing legal precision and accessibility for diverse types of claims. One notable similarity is the integration of lay judges, though their prominence varies. Denmark and Norway utilize lay judges extensively across civil and criminal proceedings. These non-professional adjudicators help democratize legal decisions, ensuring that community perspectives influence legal interpretation and verdicts. This participatory model not only reinforces public trust but also distributes judicial authority beyond the elite legal class. Sweden, by contrast, limits the use of lay judges primarily to criminal cases, relying more heavily on professional judges. This approach may enhance consistency but risks distancing legal outcomes from public sentiment. The philosophical contrast—between technocratic adjudication and participatory justice—reflects deeper national views on the role of civic engagement in state institutions.

Legal aid structures also reveal significant variation in equity and access. Denmark's model is notably inclusive: legal aid is widely accessible, supported by both a robust public defender system and legal literacy initiatives. Norway legally guarantees aid, though accessibility may depend on procedural thresholds and state funding priorities. Sweden applies a means-tested system with a sliding scale for cost-sharing, which can leave some individuals just above eligibility limits without support. These frameworks contrast starkly with the United States, where legal aid is highly fragmented and underfunded, often failing to meet the needs of indigent defendants and civil litigants alike. The Scandinavian models, by contrast, demonstrate how legal infrastructure and funding choices directly influence the equity of legal outcomes. Finally, these countries' broader socio-economic conditions and legal philosophies shape how justice is experienced across different segments of society. Denmark's low Gini coefficient (30.0 in 2022) and its commitment to human rights—through institutions like the Danish Institute for Human Rights—create a legal ecosystem attuned to both procedural fairness and social equity. Norway's exceptionally low Gini coefficient (27.6) and its justice system's emphasis on rehabilitation (reflected in its 25% recidivism rate) further highlight the alignment between social policy and judicial philosophy. Sweden reinforces its egalitarian commitments through the Discrimination Act and the work of the Equality Ombudsman, which together institutionalize protections for disadvantaged populations. While all three countries achieve relatively high levels of equity, their systems face limitations: mediation-heavy processes may pressure vulnerable litigants into unfavorable settlements (in Denmark), legal aid may be resource-constrained (in Norway), and means-testing may exclude the near-poor (in Sweden). These distinctions illuminate how institutional design—not merely intent—shapes the success of justice systems in combating inequality.

Although there is no singular blueprint for a successful democracy, democratic legitimacy can be realized through different institutional paths, each shaped by history, culture, and values. While Denmark and Norway emphasize participatory justice through lay judges and broad legal aid access, Sweden prioritizes professional adjudication alongside transparency and anti-discrimination oversight. These differences do not undermine their democratic integrity;

instead, they reflect the adaptability of democratic principles across contexts. Crucially, what unites these models is not uniformity, but their shared commitment to fairness, inclusion, and accountability. This comparative perspective matters deeply for the United States, where high income inequality, limited legal aid, and adversarial procedures have entrenched barriers to justice. Although the U.S. differs structurally and culturally, it is not incapable of democratic renewal. Reforms such as expanding civil legal aid, introducing participatory mechanisms like lay judges or community courts, and decoupling legal outcomes from wealth could realign American justice with democratic ideals. The Scandinavian examples offer not a rigid template, but a provocation: that justice systems, when deliberately designed, can be both egalitarian and effective—and that the U.S., too, can build a democracy that is not only representative in theory, but substantively fair in practice.

V. Suggestions

In contrast to the nation's current punitive system, the overarching suggestion is to embrace restorative justice principles, a framework that has guided several European countries. To emphasize these values, this section highlights the author's suggestions for the United States; these are tangible, utilitarian recommendations in an ambitious bid to annihilate wealth, racial, and socioeconomic inequality in the country.

i. Reduce Economic Barriers: *Ensure that individual wealth plays no role in law.*

Nationalize or Standardize Civil Legal Aid: Guarantee cost-free and adequate legal resources not only in criminal defense, but also in essential civil domains of housing, family law, or immigration, for outcomes that can be equally life-altering (as established in *Gideon v. Wainwright*).

Sliding Scale Legal Aid: Adopt Sweden's means-tested model to create a tiered support system, ensuring that individuals slightly above the poverty line are not excluded from assistance, while preserving equitable access across income levels.

Fund Public Defender Offices More Robustly: Substantially increase funding and staffing for public defender offices to combat chronic under-resourcing and excessive caseloads that result in disparities in the quality of defense provided.

ii. Build Public Trust: *Promote Judicial Transparency and Accountability.*

Revise judicial appointment systems: Increase transparency, reduce partisanship and establish nonpartisan selection panels, similar to those of Denmark's Judicial Appointments Council, to ensure that judges are chosen for competence and integrity rather than political alignment.

Civic Education: Expand accessible civic education programs that inform citizens of their legal rights and the justice process, as demonstrated by Scandinavia's participatory legal culture. Additionally, form citizen review boards or community forums to assess court responsiveness regarding equity and inclusion.

Introduce Lay Judges: Implement pilot lay judge programs, beginning with local courts (e.g., housing or family court); involve trained citizens alongside professionals, particularly in regions with high public distrust in the system.

Expand community courts: Use community dispute resolution models to mirror Denmark's emphasis on mediation while ensuring representation safeguards. Additionally, support more open proceedings, like that of Sweden's oral hearings and public trial emphasis, ensuring court processes are transparent and accessible to the public.

iii. Consider Alternative Methods for Youth and Low-Level Offenders: *Encourage Restorative Journeys for Offenders.*

Diversion Programs with Restorative Options: Drawing from Norway's restorative justice model, implement diversion programs emphasizing rehabilitation and personal accountability by reducing recidivism and humanizing the justice process.

Victim-offender reconciliation programs: Establish structures and voluntary dialogue spaces between victims and offenders, facilitated by trained mediators under legal supervision, aiming to foster understanding and strengthen community cohesion.

iv. Create specialized courts: *Recognize the Uniqueness of Each Case.*

Establish More Issue-Specific Courts: Borrowing from Denmark and Norway's use of administrative and land courts, establish more issue-specific courts (e.g., environmental, labor) to improve legal accessibility.

Together, these reforms effectively target multidimensional barriers that inhibit equal access to economic and institutional justice. Expanding legal aid and funding defense infrastructures will dismantle the economic stratification that currently conditions individual's capacity to claim legal rights. Simultaneously, embedding transparency in legal processes through reformed appointment processes and participatory mechanisms such as lay judges or community courts will recalibrate the judiciary's relationship with the public in an ambitious bid to transform from a distant authority into a responsive civic institution. Meanwhile, restorative programs and specialized courts would reorient legal practice toward contextualized, rehabilitative, and expertise-driven adjudication, allowing acknowledgement of nonmonolithic justice, such as relational and situational justice. In their totality, these measures will allow justice to become a shared public good and social trust, rather than its initial suspicion of a privilege contingent upon wealth or proximity to power. The following section will synthesize these findings within the theoretical overview of procedural justice and institutional legitimacy, situating these proposed models with comparative global discourses.

VI. Conclusion

In this analysis, we have examined the multifaceted ways in which wealth continues to shape legal outcomes in the United States, undermining the foundational democratic ideals of liberty, equality, and justice that are theoretically enshrined in the Constitution. We first looked

at historical and philosophical frameworks, particularly those of Plato and Aristotle's deliberations on equity and impartiality. These Enlightenment principles became embedded into modern constitutional designs, establishing the aspiration that legal systems should operate independent of external social and economic pressures. Still, intergenerational inequalities and structural deficiencies in education and employment perpetuate disparate access to justice, constraining socioeconomic mobility for low-income individuals. With this disparity, tangible disadvantages in both civil and criminal proceedings hinder low-income communities from gaining fair chances under the law. To worsen, systemic corruption and discretionary power exercised within law enforcement and judicial decision-making, which inhibit implicit biases and resources asymmetries, create a two-tiered justice system that privileges the elites.

Secondly, we conducted comparative analyses of Scandinavian systems: Denmark, Norway, and Sweden's alternative approaches that materially mitigate the influence of socioeconomic status on legal outcomes. Clearly, these jurisdictions are effective because of their robust legal aid infrastructures, participatory mechanisms, and restorative justice principles. For America, we can obtain equitable access and reduce economic barriers and support the excluded by designing our legal aid systems closely to those of Scandinavian countries'.

With these insights, it becomes clear that the United States' persistent inequities are structurally embedded rather than merely incidental. Reform strategies must therefore be multidimensional: expanding civil and criminal legal aid, support systems, and increasing funding for public defenders. These directly address the economic barriers that inhibit legal participation. More specifically, reforms to judicial appointment processes, civic education initiatives, and the introduction of participation mechanisms, like lay judges or community courts, recalibrate the relationship between the judiciary and the public. Furthermore, the development of restorative programs and specialized courts situate adjudication within a context-sensitive framework that recognizes the relational and situational nature of justice. Collectively, these interventions target the structural dimensions of inequity, creating a legal system in which access to justice is not contingent upon wealth, but instead treated as a public good undergirded by social trust and procedural fairness.

Through deliberate policy design informed by comparative models, the United States can aspire to fulfill the promise of justice that has long been foundational to its democratic project.

Works Cited

- “About the Judges Proposals Board.” Domarnämnden, 2021, www.domstol.se/en/domarnamnden/information-in-english/.
- Aristotle. *The Nicomachean Ethics*. Harvard University Press, 1934.
- Bound, John, et al. “Why Have College Completion Rates Declined? An Analysis of Changing Student Preparation and Collegiate Resources.” *American Economic Journal: Applied Economics*, vol. 2, no. 3, July 2010, pp. 129–57, <https://doi.org/10.1257/app.2.3.129>.
- Cantone, Jason A., et al. “Sounding Guilty: How Accent Bias Affects Juror Judgments of Culpability.” *Journal of Ethnicity in Criminal Justice*, vol. 17, no. 3, June 2019, pp. 228–53, <https://doi.org/10.1080/15377938.2019.1623963>.
- Cantor, Henry. “Law Office of Arash Hashemi.” Los Angeles Criminal and Self Defense Lawyer, 7 Oct. 2024, www.hashemilaw.com/cost-of-felony-defense-lawyer-california/. Accessed 22 Oct. 2025.
- Constitution Annotated. “Civil War Amendments (Thirteenth, Fourteenth, and Fifteenth Amendments) | Constitution Annotated | Congress.gov | Library of Congress.” Congress.gov, Library of Congress, 2024, constitution.congress.gov/browse/essay/intro.6-4/ALDE_00000388/.
- Danish. “Evaluation of Danish Support to Promotion and Protection of Human Rights 2006-2016.” UM-ENEN, 2016, um.dk/en/danida/results/eval/eval_reports/human-rights-20180619t142800.
- Dee Gill. “Smartphone Records Reveal Racial Disparities in Neighborhood Policing.” *UCLA Anderson Review*, 25 Oct. 2023, anderson-review.ucla.edu/smartphone-records-reveal-racial-disparities-in-neighborhood-policing.
- “Denmark - Court Administration - EJTN.” European Judicial Training Network, 21 Aug. 2024, ejtn.eu/office/denmark-court-administration/. Accessed 22 Oct. 2025.
- “Denmark | Judiciaries Worldwide.” *Judiciariesworldwide.fjc.gov*, judiciariesworldwide.fjc.gov/country-profile/denmark.
- Diesen, Christian. “The Advantages and Disadvantages of Lay Judges from a Swedish Perspective.” *Revue Internationale de Droit Pénal*, vol. 72, no. 1, 2001, p. 355, <https://doi.org/10.3917/ridp.721.0355>.
- “Discrimination Act, 2008:567.” *Www.do.se*, 2025, www.do.se/choose-language/english/discrimination-act-2008567.
- Egede, Leonard E., et al. “Modern Day Consequences of Historic Redlining: Finding a Path Forward.” *Journal of General Internal Medicine*, vol. 38, no. 6, Feb. 2023, pp. 1534–37, <https://doi.org/10.1007/s11606-023-08051-4>. PubMed Central.
- Emma Zack. “Why Holding Prosecutors Accountable Is so Difficult.” Innocence Project, 23 Apr. 2020, innocenceproject.org/news/why-holding-prosecutors-accountable-is-so-difficult/.
- Ghandnoosh, Nazgol. “One in Five: Disparities in Crime and Policing.” The Sentencing Project, The Sentencing Project, 2 Nov. 2023, www.sentencingproject.org/reports/one-in-five-disparities-in-crime-and-policing/.

- Glynn, Jennifer. "Opening Doors: How Selective Colleges and Universities Are Expanding Access for High-Achieving, Low-Income Students." Jack Kent Cooke Foundation, Sept. 2017, www.jkcf.org/research/opening-doors-how-selective-colleges-and-universities-are-expanding-access-for-high-achieving-low-income-students/.
- Goldbach, Toby, and Valerie Hans. "Juries, Lay Judges, and Trials." Cornell Law Faculty Working Papers, vol. 122, Jan. 2014, scholarship.law.cornell.edu/clsoops_papers/122/.
- Goldberg, Deborah. Public Funding of Judicial Elections: Financing Campaigns for Fair and Impartial Courts. www.prisonpolicy.org/scans/jiseries3.pdf. Accessed 22 Oct. 2025.
- Goodman, Nanette, et al. FINANCIAL INEQUALITY: Disability, Race and Poverty in America. National Disability Institute, Feb. 2019, www.nationaldisabilityinstitute.org/wp-content/uploads/2019/02/disability-race-poverty-in-america.pdf.
- Harlow, Caroline. Defense Counsel in Criminal Cases. Nov. 2000, bjs.ojp.gov/content/pub/pdf/dccc.pdf.
- Hazard, John, and Maurice Millner. "Roman Law - the Law of Justinian | Britannica." Encyclopædia Britannica, 2019, www.britannica.com/topic/Roman-law/The-law-of-Justinian.
- HLR. "Stacked: Where Criminal Charge Stacking Happens — and Where It Doesn't." Harvard Law Review, 10 Mar. 2023, harvardlawreview.org/print/vol-136/stacked-where-criminal-charge-stacking-happens-and-where-it-doesnt/
- Högborg, Björn, et al. "Associations between Academic Achievement and Internalizing Disorders in Swedish Students Aged 16 Years between 1990 and 2018." European Child & Adolescent Psychiatry, vol. 34, no. 5, Oct. 2024, <https://doi.org/10.1007/s00787-024-02597-2>.
- "Income Inequality." Wwv.dst.dk, www.dst.dk/en/Statistik/emner/arbejde-og-indkomst/indkomst-og-loen/indkomstulighed.
- "Independence of the Courts." Domstol.no, 2022, www.domstol.no/en/about-the-legal-system/independence-of-the-courts/.
- Johnson, Allan Chester. "Ancient Roman Statutes : Translation, with Introduction, Commentary, Glossary, and Index." Yale.edu, Austin : University of Texas Press, 1961, 2019, avalon.law.yale.edu/ancient/twelve_tables.asp.
- Kandel, Sean. "An Evidence-Based Look at the Effects of Diet on Health." Cureus, vol. 11, no. 5, May 2019, <https://doi.org/10.7759/cureus.4715>.
- Kang, Michael. "Campaign Cash and Judicial Outcomes." State Court Report, 2024, statecourtreport.org/our-work/analysis-opinion/campaign-cash-and-judicial-outcomes.
- Kingsley, Eric. "Wage Theft in California: Your Rights Explained." Kingsleykingsley.com, 2024, www.kingsleykingsley.com/california-wage-theft-rights. Accessed 22 Oct. 2025.
- Lafayette, Marquis. "Declaration of Human and Civic Rights of 26 August 1789 | Conseil Constitutionnel." Wwv.conseil-Constitutionnel.fr, 1789, www.conseil-constitutionnel.fr/en/declaration-of-human-and-civic-rights-of-26-august-1789.
- "Lay Judge." Domstol.no, 2022, www.domstol.no/en/lay-judge/.

- “Legal Aid.” County Governor - Statsforvalteren.no, 2020,
www.statsforvalteren.no/en/portal/people-and-society/legal-aid/. Accessed 22 Oct. 2025.
- Levenson, Laurie. “High-Profile Prosecutors &(And) High-Profile Conflicts High-Profile Prosecutors &(And) High-Profile Conflicts.” *Oyola of Los Angeles Law Review Loyola of Los Angeles Law Review*, vol. 39, no. 4, 2006. Symposium: Celebrity Prosecutions .
- Lowery, George. “Study Uncovers Why Jurors Reward the Good-Looking, Penalize the Unbeautiful.” *Cornell Chronicle*, 11 May 2010,
news.cornell.edu/stories/2010/05/unattractive-people-pay-price-court.
- Mariotti, Agnese. “The Effects of Chronic Stress on Health: New Insights into the Molecular Mechanisms of Brain–Body Communication.” *Future Science OA*, vol. 1, no. 3, Nov. 2015, <https://doi.org/10.4155/fso.15.21>.
- Martínez, Madeline. “The Linkage between Productivity and Language Proficiency.” *Languagetesting.com*, 29 Aug. 2023,
www.languagetesting.com/blog/the-linkage-between-productivity-and-language-proficiency/.
- Maslow, Abraham. “A Theory of Human Motivation.” *Psychological Review*, vol. 50, *Psychological Review*, 1943, pp. 370–96, psychclassics.yorku.ca/Maslow/motivation.htm.
- McEwen, Bruce S. “Neurobiological and Systemic Effects of Chronic Stress.” *Chronic Stress*, vol. 1, no. 1, Apr. 2017, p. 247054701769232, <https://doi.org/10.1177/2470547017692328>.
- Morgan, Kate. “Why In-Person Workers May Be More Likely to Get Promoted.” *BBC*, 8 Mar. 2021, www.bbc.com/worklife/article/20210305-why-in-person-workers-may-be-more-likely-to-get-promoted.
- National Archives. “19th Amendment to the U.S. Constitution: Women’s Right to Vote (1920).” National Archives, National Archives, 8 Feb. 2022,
www.archives.gov/milestone-documents/19th-amendment.
- National Constitution Center. “The 0th Article of the U.S. Constitution.” National Constitution Center – *Constitutioncenter.org*, 2025, constitutioncenter.org/the-constitution/preamble.
- . “The Amendments.” National Constitution Center, 1791,
constitutioncenter.org/the-constitution/amendments.
- Neitz, Michele Benedetto. “Socioeconomic Bias in the Judiciary.” *Papers.ssrn.com*, 19 Sept. 2012, papers.ssrn.com/sol3/papers.cfm?abstract_id=2149311.
- NPR. “How Lawyers with High-Profile Clients Approach Jury Selection.” NPR, 14 Apr. 2024,
www.npr.org/2024/04/14/1244728189/how-lawyers-with-high-profile-clients-approach-jury-selection.
- Ojeda, Christopher. “Depression and Political Participation*.” *Social Science Quarterly*, vol. 96, no. 5, June 2015, pp. 1226–43, <https://doi.org/10.1111/ssqu.12173>.
- Oyez. “Gideon v. Wainwright.” Oyez, 2018, www.oyez.org/cases/1962/155.
- Petrus, Johannes, et al. LISINDIA DORELL BAILEY Fifth Applicant FUNDISWA VIRGINIA BIKITSHA Sixth Applicant MERLE BRUINTJIES Seventh Applicant. 2016,
www.saflii.org/za/cases/ZACC/2016/32.pdf.
- Plato. *The Republic*. Translated by C J Rowe, Penguin, 2012.
- Rachlinski, Jeffrey, et al. “Getting Explicit about Implicit Bias | Judicature.” *Judicature.duke.edu*, 2 Dec. 2020, judicature.duke.edu/articles/getting-explicit-about-implicit-bias/.

Regeringskansliet, Regeringen och. "The Equality Ombudsman (DO)." Regeringskansliet, 6 Mar. 2015, www.government.se/government-agencies/equality-ombudsman-do/.

Religion & Race. "What Is Internalized Oppression?" R-Squared, 2025, www.r2hub.org/library/what-is-internalized-oppression.

Schoultz, Isabel. "Legal Aid in Sweden." Springer EBooks, Springer Nature, Dec. 2017, pp. 43–76, https://doi.org/10.1007/978-3-319-46684-2_3.

South African Government. "Constitution of the Republic of South Africa, 1996 | South African Government." Wwww.gov.za, 1996, www.gov.za/documents/constitution/constitution-republic-south-africa-1996-04-feb-1997

Stinson, Jennifer Kirsten. "Police Integrity Lost: A Study of Law Enforcement Officers Arrested." Middle West Review, vol. 2, no. 2, 2016, pp. 53–84, <https://doi.org/10.1353/mwr.2016.0009>.

"Structure of the Courts." Advokatforeningen, www.advokatforeningen.no/en/features-of-the-norwegian-legal-system/structure-of-the-courts/.

Tilsley, Alexandra. "School Funding: Do Poor Kids Get Their Fair Share?" Urban Institute, May 2017, apps.urban.org/features/school-funding-do-poor-kids-get-fair-share/.

Titia Loenen. The Conceptualization of Equality and Non-Discrimination as Legal Standards. BRILL, 2025.

United Nations. "International Covenant on Civil and Political Rights." OHCHR, United Nations, 16 Dec. 1966, www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights.

---. "Universal Declaration of Human Rights." United Nations, 10 Dec. 1948, www.un.org/en/about-us/universal-declaration-of-human-rights.

Webster, Julianne. "Victimology in Context: Examining Marginalized Communities and Crime Impact." Sociology and Criminology, vol. 12, no. 2, 2025, <https://doi.org/10.35248/2375-4435.24.12.318>.

Wiser, Rooke. "How the Wealthy Delay and Derail Legal Accountability through Motion Abuse." Tower, 13 Nov. 2023, tower.mastersny.org/13429/opinion/how-the-wealthy-delay-and-derail-legal-accountability-through-motion-abuse/.

World Bank. "GINI Index (World Bank Estimate) - Norway | Data." Worldbank.org, 2011, data.worldbank.org/indicator/SI.POV.GINI?locations=NO.

World Justice Project. "WJP Rule of Law Index 2023 Global Press Release." World Justice Project, 25 Oct. 2023, worldjusticeproject.org/news/wjp-rule-law-index-2023-global-press-release.

Yohnka, Ed, et al. "Ending Qualified Immunity Once and for All Is the next Step in Holding Police Accountable | News & Commentary." American Civil Liberties Union, 23 Mar. 2021, www.aclu.org/news/criminal-law-reform/ending-qualified-immunity-once-and-for-all-is-the-next-step-in-holding-police-accountable.

Relationship Between Parental Education and Financial Literacy of Youth in India By Advay Gupta

Abstract

This study investigates the relationship between parental education and the financial literacy of adolescents in India. Data were collected through a structured questionnaire administered across multiple schools and colleges. Using data from 171 respondents aged 13–18 years, the research explores how parental educational attainment, age, gender, and learning sources influence financial knowledge, attitudes, and behaviors. Correlation analysis reveals no significant relationship between either mother's or father's education and the financial literacy of their children ($r = 0.011$ and $r = -0.012$, respectively). Age emerged as the strongest predictor ($r = 0.17$), indicating that financial literacy modestly increases with maturity, while gender differences were negligible. Notably, respondents who identified the internet as their primary learning source achieved higher average literacy scores than those learning from parents or schools. The findings suggest that financial competence among adolescents is increasingly independent of parental background and shaped by self-directed, technology-based learning. The study proposes policy recommendations emphasizing digital financial education, experiential and behavioral learning, and inclusive national curricula to improve financial capability among India's youth.

1. Introduction

Financial literacy is an essential life skill. Financial literacy is defined as a combination of knowledge, skills, attitudes, and behaviors needed to make effective decisions and, as a result, achieve personal financial well-being (Faulkner and Ashley, 2015). Financial literacy covers subjects like investments, insurance, funds, pensions, real estate, and tax planning, as well as more complicated concepts such as the time value of money, card operations, and compound interest. Financial literacy enables an individual to understand their fundamental needs, wants, and desires, which improves their understanding of their financial situation and helps them manage their finances more effectively (Wahab et. al., 2024).

A study (Jayaraman & Jambunathan, 2018) measuring financial literacy levels among high school students in India found low levels of performance on standard measures of financial literacy. Despite having high levels of numeracy, they were unable to transfer that knowledge to do financial computations. Interviews with students highlighted the fact that understanding of societal and macroeconomic impacts of financial literacy was low. Researchers have found that young adults are less likely to think about retirement (van Rooij et al 2011), participate in the stock market (van Rooij et al 2011), and save for emergencies (Babiarz & Robb 2014; Zumbrun 2014). Young adults are also significantly more likely to be underbanked (Breitbach & Walstad 2014). All of these studies suggest that young adults are more likely to engage in financial

behaviors and practices that could be costly and have long-term adverse consequences on their financial well-being.

With Gen-Z being the newest entrants in the workplace, financial literacy becomes even more significant for them. Poor financial literacy has been shown to affect workplace efficiency and yield (Garman et al, 1996, 1999; Joo and Grable, 2000; Kim et al., 1998; Kim and Garman, 2004). In India, where financial literacy among youth remains alarmingly low despite growing access to banking and digital finance tools, there is a need to understand the factors that influence financial knowledge and behavior. This study addresses that gap by focusing on the relationship between parental educational attainment and the financial literacy of Indian youth. Most existing studies on financial literacy analyze broader factors like income or urban/rural background, but not parental education as an independent variable.

This research attempts to examine whether and to what extent the educational attainment of parents influences the financial literacy levels of their children. It investigates if parental education serves as a direct contributor to financial knowledge of the youth or does it affect the youth's attitude towards financial situations. Theoretically, parental education may influence financial literacy through knowledge transfer and modeling of financially responsible behaviors (Lusardi et. al (2010)). The research also seeks to explore whether these effects differ across various levels of parental education. If parental education is found to have a strong influence, then youth from households with lower parental education may be given additional support to close the financial knowledge gap. This research contributes to the limited body of work in India that examines family background as a determinant of financial literacy. The findings of this study can help shape more effective and inclusive financial education policies.

2. Literature Review

2.1 Financial literacy among adolescents

Existing literature consistently reports that young adults tend to have lower levels of financial literacy compared to older cohorts (Lusardi & Mitchell, 2014; Allgood & Walstad, 2013). One likely reason for this gap is a lack of experience in making financial decisions. Agarwal et al. (2009) suggest that younger adults are more prone to mistakes, including paying higher interest rates and incurring costly credit card fees. This vulnerability is further reflected in their limited engagement with long-term financial planning: studies indicate that young adults are less likely to save for retirement, participate in the stock market (Rooij et al. 2011), or maintain emergency savings (Babiarz & Robb, 2014; Zumbun, 2014). Additionally, they are more likely to be underbanked, which may limit access to financial tools and resources (Breitbach & Walstad, 2014).

Credit card usage provides a concrete example of how these literacy gaps are seen in everyday behavior. Robb et al. (2009) found that college students with higher financial literacy were more likely to use credit cards responsibly, whereas Ludlum et al. (2012) revealed that the majority of students lacked basic knowledge about interest rates, late fees, and other common

charges. Collectively, these findings highlight that low financial literacy in young adults not only limits their ability to plan effectively but also increases the likelihood of costly financial mistakes.

2.2 Parents and financial literacy

Several studies show how parents influence the financial literacy of their children. Nano and Istrofor (2017) argued that parental income is one factor that can affect a student's financial literacy level. Rahman and Wahab (2023) emphasized that parents impact their children's thinking, and this influence can be mutual. Parents who lack financial literacy themselves are less able to provide relevant financial guidance to their children. In contrast, individuals from higher-income families tend to have stronger financial literacy, because parents with greater resources are more likely to provide their children with pocket money and encourage responsible money management.

For example, a study conducted in Albania (Nano et al. 2017) found that parental income significantly affected the financial literacy of both male and female students. Oseifuah et al. (2018) explained that parents with higher incomes often give children more pocket money and assign them responsibility over their own finances. This responsibility motivates children to learn financial management from their parents and increases their use of financial services. Similarly, Lusardi et al. (2010) found that financial literacy was positively correlated with parental education and whether parents held stocks or retirement accounts while their children were growing up. Collectively, these studies suggest that parental influence operates through multiple mechanisms, including knowledge transfer, role modeling, and providing opportunities for practical financial experience.

2.3 School and financial literacy

Researchers have examined the role of schools in shaping financial literacy (Choi (2010), Danes and Brewton (2010), Ryack (2011) etc.). Data from the Jump\$tart Coalition Survey of Personal Financial Literacy among students, which includes 31 questions designed to measure financial knowledge, largely indicate no significant relationship between formal financial education and literacy levels. Specifically, students who participated in dedicated courses on money management, personal finance, or economics did not score significantly higher than those who had no such courses. However, Shim, Xiao, et al. (2009) and Shim, Barber, et al. (2010) used surveys of college students to explore the relationship between financial knowledge and well-being. Their analyses, including bivariate and multiple equation models, found that financial education in high school was correlated with self-assessed financial knowledge in college. These findings suggest that the effectiveness of school-based financial education may depend on students' personal interest, motivation, or engagement with the subject.

While previous studies (Herman et al., Mandell and Klein (2007), and Choi (2010)) have examined the role of formal education and income in shaping financial literacy, the impact of parental education remains underexplored and little is known about whether these relationships

hold true in the Indian context, where family structures and education systems differ substantially.

Thus, while institutional education contributes modestly, this study focuses on the parental dimension, an area still underexplored in Indian research.

3. Research Methodology

This research is specifically designed to investigate the relationship between parental education and financial literacy levels among adolescent youth in India. It will be conducted across various schools and colleges throughout the country. To measure financial literacy, the questionnaire includes items covering financial knowledge, financial attitude, and financial skills. The remainder of the questionnaire collects information on parents' education, occupation, and income, along with respondents' academic stream and their perception of how financially literate they are. This perceived literacy is compared to their actual scores. The financial literacy scores are then correlated with parental education levels to determine whether a relationship exists between the two variables.

3.1 Data Collection

The target population consists of participants aged 13–18 years. Data was collected through a structured questionnaire administered using Google Forms. This Google form platform supports various question formats (e.g., multiple-choice, Likert scales), allowing for clearly defined, quantifiable measures. Its ease of deployment, flexibility in question design, and ability to efficiently gather data across diverse demographics make it a suitable instrument for this study (Ann & Chinnaswamy, Brindhadevi, 2025). However, reliance on Google Forms restricts participation to those with internet access, potentially excluding marginalized groups (Couper, 2000).

Participation was voluntary, and respondents provided informed consent prior to completing the survey.

3.2 Sample Size

The study employs a sample size of 171 respondents. This number was chosen as it balances feasibility and statistical significance. Previous studies assessing youth financial literacy have used similar sample sizes, such as Jayaraman & Jambunathan (2018) in India and Robb & Sharpe (2009) in the United States. Furthermore, according to the Central Limit Theorem, a sufficiently large sample ($n \geq 30$) approximates a normal distribution. With 171 participants, the dataset ensures greater reliability, reduced sampling error, and adequate power for inferential statistical tests. At this size, the margin of error remains within acceptable limits (approximately $\pm 7\%$ at a 95% confidence level), making 200 respondents a justifiable and statistically valid choice (Cochran, 1977; Krejcie & Morgan, 1970).

3.3 Variables

The study focuses on two primary variables. The independent variable is parental education, measured by the highest level of formal education attained by either parent. The dependent variable is the financial literacy of young adults, assessed through the questionnaire, which combines knowledge-based questions with items on attitudes and behaviors. By positioning parental education as the predictor and financial literacy as the outcome, the study seeks to determine the extent to which parents' educational attainment influences their children's financial capabilities.

3.4 Data Analysis

A quantitative approach is used to interpret the data and assess the relationship between these two variables. Quantitative methods provide important strengths: structured surveys allow for systematic data collection from 171 respondents, ensuring comparability and statistical reliability (Creswell, 2014). By including both multiple-choice and Likert-scale questions, the survey captures not only factual knowledge but also financial attitudes and perceptions, offering a broader view of financial literacy (Lusardi & Mitchell, 2011). Multiple-choice items are effective for objectively measuring knowledge of concepts such as budgeting, interest, and inflation in a standardized format, ensuring consistency across responses. In contrast, Likert-scale questions capture attitudes, behaviors, and confidence levels by allowing respondents to express degrees of agreement or disagreement, thus providing more nuanced insights into financial practices (Xiao, Chen & Chen, 2014). Moreover, quantitative methods minimize researcher bias and enable objective analysis when evaluating the connection between parental education and financial literacy (Neuman, 2014).

3.5 Limitations

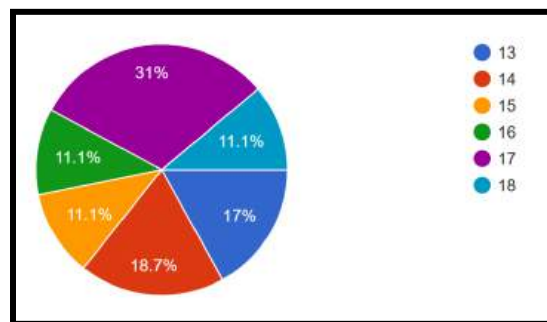
While this study offers valuable insights, several limitations should be acknowledged. First, the questionnaire includes only six financial literacy questions, which may not fully capture the breadth of knowledge, attitudes, and behaviors that constitute financial literacy (Hung, Parker & Yoong, 2009). Although the instrument is described as "scientifically validated," the absence of a clear reference to its validation study reduces certainty regarding its reliability. Attitude-based items also carry the risk of social desirability bias, as respondents may provide answers they believe to be more acceptable rather than their true opinions. Furthermore, the categorical measurement of parental education may limit precision and comparability across different educational levels. The use of Google Forms further introduces bias toward respondents with reliable internet access and digital literacy, which may underrepresent rural or low-income populations (Etikan et al., 2016). Finally, the exclusive use of quantitative interpretation restricts the ability to capture the depth of personal financial experiences that qualitative data might reveal (Bryman, 2016). Relying solely on quantitative interpretation excludes deeper personal insights that qualitative approaches could capture (Bryman, 2016). The narrow scope of variables, focusing mainly on parental education, overlooks additional influences such as peers or exposure to financial education programs (Atkinson & Messy, 2012).

4. Data Analysis

The data analysis is organized into two components. The first involves a question-by-question examination of survey responses. For each item, the overall distribution of answers is summarized, illustrated through graphical representations, and interpreted. The second component consists of correlation analyses, undertaken to explore the relationships between financial literacy and parental education. These patterns provide a context for understanding how family and demographic backgrounds may (or may not) influence financial literacy.

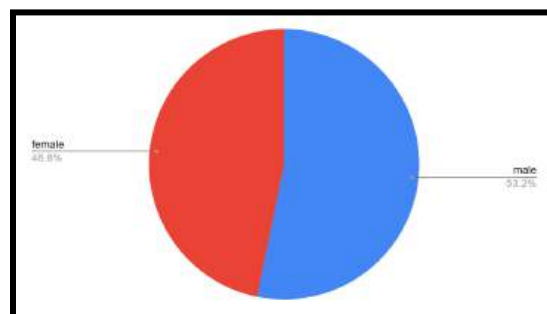
Demographics

4.1 Age Distribution



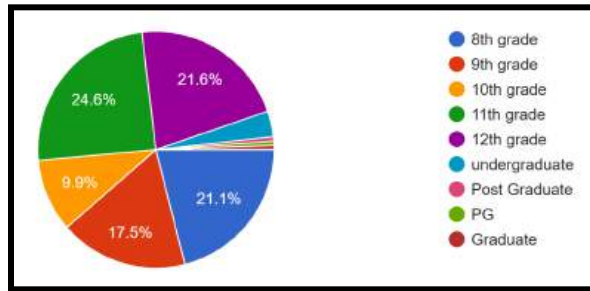
The age distribution of respondents is generally balanced, with the largest proportion being 17-year-olds (31%). The proportions of 13–14 year-olds are nearly equal, while the shares of 15, 16, and 18-year-olds are identical at 11.1% each. The distribution ensures representation across adolescence, allowing for meaningful comparison of financial literacy across age groups.

4.2 Gender Distribution



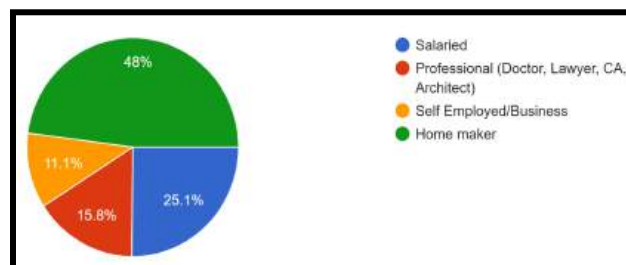
The sample consisted of 53.2% male and 46.8% female respondents, indicating a nearly balanced distribution across gender, with a slight predominance of male participants and diverse academic specialisation.

4.3 Level of Education of Respondents



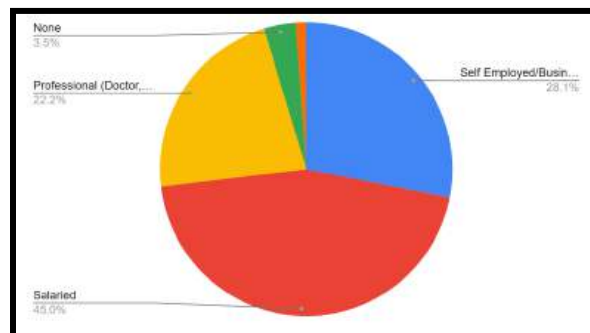
The highest education level of respondents is concentrated in secondary school, with 11th grade (24.6%) forming the largest group, followed by 12th grade (21.6%), 8th grade (21.1%), and 9th grade (17.5%). Only 9.9% were in 10th grade, This shows the sample is primarily secondary school students, especially in 11th and 12th grades.

4.4 Mother's Occupation



The data shows that nearly half of the respondents' mothers (48%) are homemakers, while 25.1% are salaried employees. Professionals (15.8%) and those self-employed or in business (11.1%) form smaller segments. Homemaking thus emerges as the predominant occupation.

4.5 Father's Occupation

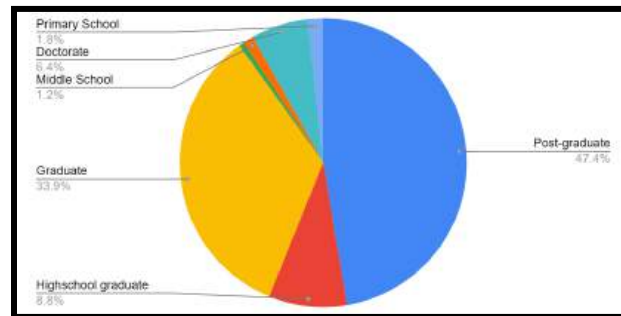


Among respondents' fathers, 45% are salaried, 28.1% are self-employed or in business,

and 22.2% work in professional roles. A small share (3.5%) reported no occupation. Salaried employment emerges as the most common, with significant representation of professionals and entrepreneurs

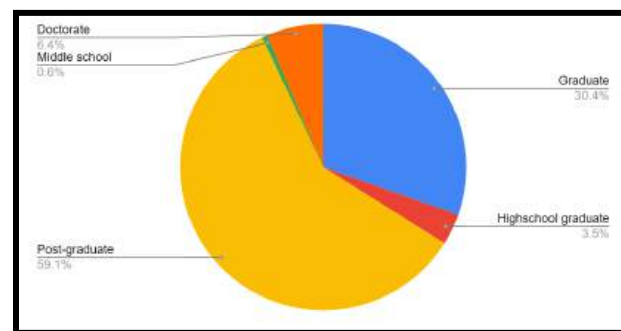
4.6 Parent Education Classification

Mother's Level of Education



The findings reveal that a substantial majority of mothers possess higher education qualifications, with 47.4% holding postgraduate degrees and 33.9% being graduates. In contrast, only a small fraction reported lower educational attainment. This reflects a predominantly well-educated background for the mothers.

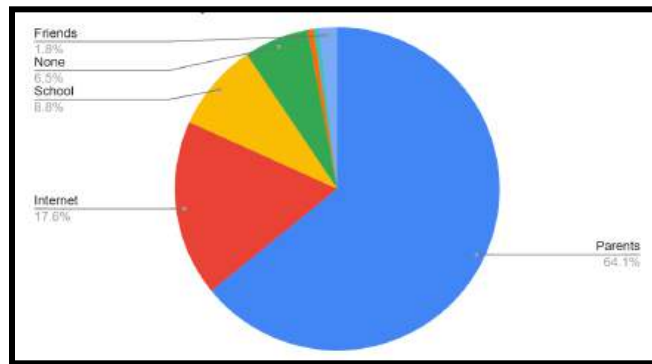
Father's Level of Education



The results indicate that fathers in the sample are predominantly highly educated, with 59.1% holding postgraduate degrees and 30.4% being graduates. Only a small proportion reported high school (3.5%) or lower levels of education. This reflects a strongly educated background for the fathers as well.

This indicates that the sample predominantly represents families with high educational attainment, which may influence exposure to financial resources.

4.7 Source of Financial Education



The data reveals that parents play the most significant role in shaping financial education, cited by 64.1% of respondents as their primary source. Other notable sources include the internet (17.6%) and schools (8.8%), while friends (1.8%) and no exposure (6.5%) accounted for minimal influence.

After examining the demographic characteristics of the sample, the following section presents the financial literacy test results and explores the extent to which these outcomes vary across demographic factors.

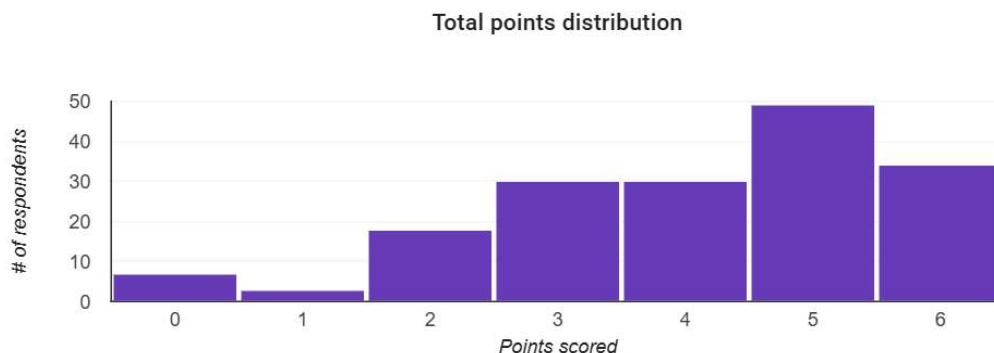
5. Results - Correlation Analysis

Financial Literacy Assessment

The financial literacy assessment consisted of six scenario-based questions, each carrying one mark, resulting in a maximum possible score of six points. The test covers compound interest, inflation effects, investment risk, mortgage calculations, insurance planning, and risk-averse investment strategies.

The correlation analysis was conducted by comparing respondents' financial literacy scores (ranging from 0 to 6) with the educational levels of their parents. Each level of education was assigned a numerical value on a scale from 0 (no schooling) to 6 (doctorate) to enable quantitative comparison. The objective was to examine whether higher parental education is associated with better financial literacy among respondents. The results of the correlation analysis are presented below, with the coefficients indicating both the strength and direction of the relationship.

Average 4.08 / 6 points	Median 4 / 6 points	Range 0 - 6 points



The financial literacy test results indicate that most respondents demonstrated a moderate level of financial understanding, with the average score being 4.08/6 points. The largest share scored 5 (49 respondents) and 6 (34 respondents), suggesting a generally competent grasp of basic financial concepts. Lower scores were less common, with 30 respondents each scoring 3 and 4, and only a small number scoring 0 (7 respondents), 1 (3 respondents), or 2 (18 respondents). Overall, the data suggests that while the majority possess satisfactory financial knowledge, a notable minority still exhibit limited financial literacy.

Mother's education and Financial literacy

The correlation between Mother's education level and financial literacy score was found to be $r = 0.011$. This indicates a weak, positive relationship. In other words, as a mother's education level increases, financial literacy tends to increase. However, the correlation is not statistically significant, suggesting that while there is some association, it should not be interpreted as a strong predictive factor.

Father's education and Financial literacy

The correlation between Father's education level and financial literacy score was found to be $r = -0.012$. This indicates a weak, negative relationship. In other words, as a father's education level increases, financial literacy tends to decrease. However, the correlation is not statistically significant, suggesting that while there is some association, it should not be interpreted as a strong predictive factor.

These findings suggest that there is no significant relationship between the parents' education and the child's financial literacy.

Source of Financial Education and Financial literacy level

The mean financial literacy score of respondents who identified their parents as their primary source of financial education was 4.08, and those whose primary source was school

achieved an average of 3.87, while those who cited the internet as their main source achieved a slightly higher mean score of 4.30. This suggests that, within the sample, respondents who rely on digital platforms for financial learning tend to exhibit marginally higher financial literacy levels compared to those primarily educated by their parents. A possible explanation for this finding is that respondents who use the internet as their primary source of financial education may be learning out of personal interest and intrinsic motivation. This aligns with the findings of Hermann et al., who noted that the effectiveness of financial education, particularly in school-based settings, often depends on students' individual interest and motivation.

Age and Financial literacy

The correlation between age and financial literacy was found to be 0.17, indicating a weak positive relationship. The correlation between age and financial literacy was found to be **0.17**, indicating a weak positive relationship. This suggests that while financial literacy tends to increase slightly with age, the association is not strong enough to imply that age alone is a significant determinant of financial literacy among adolescents. However, despite its weakness, this correlation was the strongest among all the demographic indicators analyzed, suggesting that age may still play a modest role compared to other background variables.

Gender And Financial Literacy

The analysis of gender and financial literacy revealed minimal variation between male and female respondents. The average financial literacy score for males was **4.10**, compared to **4.05** for females, indicating a negligible difference. This suggests that gender does not play a significant role in determining financial literacy levels among the surveyed youth.

Overall, the results suggest that demographic characteristics, including parental education and gender, exert minimal influence on financial literacy, while age shows a modest positive association. The data indicates that financial literacy among Indian adolescents in this sample is moderate, with limited demographic influence. The following discussion interprets these findings in the context of existing research and explores potential explanations and policy implications.

6. Discussion

6.1 Demographic Breakdown

The results of this study indicate that parental education has limited predictive value for financial literacy among Gen-Z youth in urban India. Correlation coefficients close to zero ($r = 0.011$ and $r = -0.012$) suggest no meaningful relationship between parental educational attainment and the financial knowledge of their children. This challenges the traditional assumption that higher parental education necessarily translates into better financial literacy among children (Lusardi et al. (2010); Nano et al. (2017)). According to social learning theory (Bandura, 2017), children acquire financial habits through observation and imitation. Instead,

the findings point toward the importance of individual initiative and digital exposure in shaping financial understanding, emphasizing that financial competence may be formed independent of family educational background.

The analysis further highlights the growing role of self-directed, digital learning. Respondents who relied on the internet as their primary source of financial knowledge scored higher on average (4.30) than those learning from parents (4.08) or schools (3.87). This pattern supports earlier work by Shim et al. (2010) and Xiao, Chen, and Chen (2014), who found that experiential and self-directed learning enhances financial decision-making. This trend reflects a broader shift toward digital platforms as effective tools for acquiring financial education. Consequently, financial literacy initiatives should recognize online media, interactive applications, and gamified learning as viable complements.

Age emerged as a modest but meaningful factor ($r = 0.17$), suggesting that cognitive maturity and experiential exposure contribute to financial understanding. This result aligns with Agarwal et al. (2009), who noted that financial decision-making improves with age as individuals gain experiential exposure and self-regulatory capacity. While the correlation remains weak, it supports the notion that financial education can benefit from age-specific tailoring, introducing more complex concepts progressively during later adolescence. In contrast, gender differences were negligible (males = 4.10, females = 4.05), indicating that urban Indian youth exhibit no significant gender gap in financial literacy. This finding contrasts with the international perspective (Hasler, 2017) and implies that inclusive, skill-based interventions may be more effective than gender-targeted programs in this context.

6.2 Research Implications

The study contributes to the body of literature on financial literacy by offering empirical evidence from an Indian adolescent sample, yet it also opens avenues for future research. The absence of a statistically significant link between parental education and financial literacy challenges prevailing assumptions in international literature and suggests that contextual, cultural, and technological factors may affect this relationship. Future research should therefore adopt a framework that considers socio-cultural variables alongside individual learning sources and behavioral tendencies.

Given that the present sample was predominantly urban and comprised families with high parental educational attainment, subsequent studies should expand the demographic scope to include rural regions, varied schooling systems, and lower-income groups. A more heterogeneous sample would help validate whether the observed patterns hold across India's socio-economic diversity.

Furthermore, while the current study employed a quantitative approach, future research could integrate qualitative analysis, such as interviews or focus group discussion to gain deeper insights into how adolescents understand financial concepts, and how parents, teachers, and digital media shape these learning experiences. Such mixed-method designs would enhance the explanatory power of financial literacy research in emerging economies. It would also be helpful

to examine additional predictors and variables, including peer influence, social media exposure, financial technology usage, and psychological concepts such as risk tolerance or self-efficacy. These factors may provide a more holistic understanding of how financial literacy develops in a digitally connected generation.

Lastly, researchers should aim to develop a model or measure of adolescent financial literacy tailored to the Indian context. Such a framework could integrate traditional socio-demographic determinants with new-age influences like digital engagement, behavioral orientation, and experiential exposure, to extend global financial literacy theory to reflect the reality of developing economies.

6.3 Policy Implications

The findings of this study have several implications for policymakers, educators, and financial institutions in India. The absence of a significant relationship between parental education and youth financial literacy suggests that financial competence develops largely outside the household.

First, financial literacy education should be introduced early and embedded within school curricula at both primary and secondary levels. Such programs should go beyond theory and emphasize practical application such as budgeting exercises, mock investment projects, and case-based discussions on financial decision-making. The National Education Policy (NEP) 2020 already emphasizes skill-based education, and financial literacy could be operationalized within this framework as a critical life skill.

Second, since digital platforms emerged as a more effective learning source in this study, The Government should leverage technology to deliver interactive, gamified, and accessible financial education. Collaborations with fintech and EdTech companies could help design engaging mobile applications and online modules tailored to adolescent learners. Integrating such initiatives within Digital India and National Financial Literacy Frameworks would ensure access across socio-economic groups.

Finally, given that parental education does not significantly predict financial literacy, policy approaches should adopt an inclusive and universal design rather than targeting interventions based on parental background. Special attention should be paid to marginalized and rural communities where both digital access and financial awareness remain limited. Such inclusive strategies would ensure that all young individuals, regardless of family background, have equal opportunity to acquire essential financial skills.

7. Conclusion

The findings of this study provide understanding of the factors shaping financial literacy among Indian adolescents. This study provides empirical evidence that parental educational attainment does not significantly influence financial literacy among Indian adolescents. This outcome underscores a generational shift in the sources of financial learning away from the family environment and toward independent, digital sources. The modest positive effect of age

suggests that financial competence develops gradually with cognitive maturity, while the negligible gender differences point to increasing parity in access to financial knowledge. Overall, the results imply that improving youth financial literacy requires policy and educational interventions that are systemic rather than family-dependent. Integrating practical, and technology-driven financial education into the school curriculum could strengthen financial preparedness among the next generation. The study also highlights the need for inclusive, nationwide approaches that reach adolescents across diverse socio-economic contexts.

From an academic perspective, this research contributes empirical evidence to the limited body of Indian literature on adolescent financial literacy. It calls for future studies incorporating broader samples, qualitative insights, and emerging determinants such as peer influence and fintech engagement. As India continues its rapid digital transformation, fostering financially literate youth will be critical for promoting responsible economic participation and long-term financial well-being.

Works Cited

- Agarwal, S., Driscoll, J. C., Gabaix, X., & Laibson, D. (2009). The age of reason: Financial decisions over the life cycle and implications for regulation. *Brookings Papers on Economic Activity*, 2009(2), 51–117. <https://doi.org/10.1353/eca.0.0067>
- Allgood, S., & Walstad, W. B. (2013). Financial literacy and credit card behaviors: A cross-sectional analysis by age. *Numeracy*, 6(2), 1–26. <https://doi.org/10.5038/1936-4660.6.2.3>
- Ann, M., & Chinnaswamy, V., Brindhadevi, K. (2025). Digital survey tools and their applications in education research. *Journal of Educational Methods*, 14(1), 23–32.
- Atkinson, A., & Messy, F. (2012). Measuring financial literacy: Results of the OECD / International Network on Financial Education (INFE) pilot study. *OECD Working Papers on Finance, Insurance and Private Pensions*, No. 15.
- Babiarz, P., & Robb, C. A. (2014). Financial literacy and emergency saving. *Journal of Family and Economic Issues*, 35(1), 40–50. <https://doi.org/10.1007/s10834-013-9369-9>
- Breitbart, E., & Walstad, W. B. (2014). Financial literacy and underbanked households. *Journal of Financial Counseling and Planning*, 25(1), 1–13.
- Bryman, A. (2016). *Social research methods* (5th ed.). Oxford University Press.
- Choi, L. (2010). The effect of financial education on financial literacy and financial behaviors of youth. *Journal of Family and Consumer Sciences Education*, 28(1), 63–72.
- Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). John Wiley & Sons.
- Couper, M. P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, 64(4), 464–494.
- Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). Sage Publications.
- Danes, S. M., & Brewton, K. E. (2010). The role of experiential education in learning about financial literacy. *Family and Consumer Sciences Research Journal*, 38(4), 421–435. <https://doi.org/10.1111/j.1552-3934.2010.00035.x>
- Etikan, I., Musa, S. A., & Alkassim, R. S. (2016). Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics*, 5(1), 1–4.
- Faulkner, A. E., & Ashley, D. (2015). Financial literacy: Making sense of financial education research and policy. *Journal of Economic Education*, 46(2), 189–192.
- Garman, E. T., Kim, J., Kratzer, C. Y., Brunson, B. H., & Joo, S. (1999). Workplace financial education improves personal financial wellness. *Journal of Financial Counseling and Planning*, 10(1), 79–88.
- Hasler, A. (2017). Gender differences in financial literacy: Global insights. Global Financial Literacy Excellence Center (GFLEC), George Washington University.
- Hung, A. A., Parker, A. M., & Yoong, J. K. (2009). Defining and measuring financial literacy. RAND Labor and Population Working Paper Series WR-708.
- Jayaraman, R., & Jambunathan, S. (2018). Financial literacy levels among high school students in India. *International Journal of Economics and Management Studies*, 5(7), 45–52.

- Joo, S., & Grable, J. E. (2000). Improving employee productivity: The role of financial counseling and education. *Journal of Employment Counseling*, 37(1), 2–15.
- Kim, J., & Garman, E. T. (2004). Financial stress, pay satisfaction, and workplace performance. *Compensation & Benefits Review*, 36(1), 69–76.
- Kim, J., Kratzer, C. Y., & Leech, I. E. (1998). Financial stress and absenteeism: A comparison of married and single household employees. *Journal of Family and Economic Issues*, 19(2), 161–175.
- Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. *Educational and Psychological Measurement*, 30(3), 607–610.
- Ludlum, M., Tilker, K., Ritter, D., Cowart, T., Xu, W., & Smith, B. (2012). Financial literacy and credit cards: A multi campus survey. *International Journal of Business and Social Science*, 3(7), 25–33.
- Lusardi, A., & Mitchell, O. S. (2011). Financial literacy and planning: Implications for retirement wellbeing. In O. S. Mitchell & A. Lusardi (Eds.), *Financial literacy: Implications for retirement security and the financial marketplace* (pp. 17–39). Oxford University Press.
- Lusardi, A., & Mitchell, O. S. (2014). The economic importance of financial literacy: Theory and evidence. *Journal of Economic Literature*, 52(1), 5–44.
- Mandell, L., & Klein, L. S. (2007). Motivation and financial literacy. *Financial Services Review*, 16(2), 105–116.
- Nano, R., & Istrofor, N. (2017). Parental income and student financial literacy: Evidence from Albania. *Journal of Financial Studies and Research*, 9(2), 35–46.
- Neuman, W. L. (2014). *Social research methods: Qualitative and quantitative approaches* (7th ed.). Pearson Education.
- Oseifuah, E. K., Gyekye, A., & Formadi, P. (2018). Parental income and student financial literacy: Evidence from South Africa. *African Journal of Economic and Management Studies*, 9(4), 425–438.
- Rahman, M., & Wahab, N. A. (2023). Parental influence on financial attitudes among adolescents. *Journal of Behavioral Economics and Finance*, 15(2), 113–126.
- Robb, C. A., & Sharpe, D. L. (2009). Effect of personal financial knowledge on college students' credit card behavior. *Journal of Financial Counseling and Planning*, 20(1), 25–43.
- Shim, S., Barber, B. L., Card, N. A., Xiao, J. J., & Serido, J. (2010). Financial socialization of first-year college students: The roles of parents, work, and education. *Journal of Youth and Adolescence*, 39(12), 1457–1470.
- Shim, S., Xiao, J. J., Barber, B. L., & Lyons, A. C. (2009). Pathways to life success: A conceptual model of financial well-being for young adults. *Journal of Applied Developmental Psychology*, 30(6), 708–723.
- van Rooij, M. C. J., Lusardi, A., & Alessie, R. J. M. (2011). Financial literacy and stock market participation. *Journal of Financial Economics*, 101(2), 449–472.
- Wahab, N. A., Rahman, M., & Faulkner, A. (2024). Financial literacy and personal finance behavior among youth. *International Review of Economics Education*, 45(1), 100–112.
- Xiao, J. J., Chen, C., & Chen, F. (2014). Consumer financial capability and financial satisfaction. *Social Indicators Research*, 118(1), 415–432.
- Zumbrun, J. (2014). Why are so many Americans underbanked? *The Wall Street Journal*.

Exo Ranker — A Fully Client-Side Visual-Analytics Framework for Earth-Similarity Assessment of Confirmed Exoplanets By Edward Xiong

Abstract

We present Exo-Ranker, an interactive, client-side web application for exploring and ranking confirmed exoplanets by their Earth similarity. Our workflow begins by harvesting the full “Confirmed Planets” catalogue (5934objects) from the NASA Exoplanet Archive. After cleaning and unit-standardizing six core features—planet radius (R_p), mass (M_p), equilibrium temperature (T_{eq}), and host-star effective temperature (T_s), radius (R_s), and mass (M_s)—we augment the dataset with a synthetic Earth benchmark. We then apply k-means clustering (k=4) on standardized feature vectors to reveal groups of planets with similar bulk properties, and compute an Earth Similarity Index (ESI):

Exo Ranker is an interactive, single-page web application that enables real-time exploration and quantitative ranking of the 5934 confirmed exoplanets currently listed in the NASA Exoplanet Archive. Starting from a complete snapshot of the “Confirmed Planets” table, we (i) perform rigorous cleaning and unit harmonization on six astrophysical salient features—planetary radius (R_p), planetary mass (M_p), equilibrium temperature (T_{eq}), stellar effective temperature (T_s), stellar radius (R_s), and stellar mass (M_s); (ii) append a canonical synthetic Earth record as an absolute benchmark; (iii) standardize the six-dimensional feature space and apply k-means++ clustering (k=4, chosen via the elbow criterion) to reveal coherent groups of worlds; and (iv) compute an Earth-Similarity Index (ESI) for every planet:

$$ESI = 1 - \sqrt{\frac{(R_p - 1)^2 + (M_p - 1)^2 + \left(\frac{T_{eq} - 255}{255}\right)^2 + \left(\frac{T_s - 5772}{5772}\right)^2 + (R_s - 1)^2 + (M_s - 1)^2}{6}}$$

The browser-only front end (Plotly JS) offers two linked views:

1. **Scatter view** — the user selects any pair of numeric axes; points are color-coded by cluster.
2. **Ranking view** — a live horizontal bar chart of the Top-10 planets by ESI.

To enhance accessibility, we issue a single batched call to **GPT-4o** that generates a 150-word lay summary and a coarse temperature class (Hot / Warm / Cold) for every planet; the textual output is cached in a static JSON so that the published site remains 100 % client-side with zero run-time secrets. By unifying unsupervised learning, similarity metrics, and natural-language generation, *Exo Ranker* bridges data science and science communication, helping researchers and the public alike navigate the rapidly growing exoplanet census.

1. Introduction

Since the landmark detection of 51 Pegasi b in 1995, the exoplanet inventory has expanded to thousands of worlds exhibiting extraordinary diversity in composition, orbit, and irradiation. While archival tables are indispensable to specialists, they are often opaque to non-experts and cumbersome for quick hypothesis generation. *Exo Ranker* addresses this shortfall by providing:

- Real-time dimensionality exploration of the NASA catalogue in a conventional web browser.
- A physically motivated Earth-Similarity Index that allows rapid triage of potentially habitable candidates.
- Narrative planet cards automatically distilled from raw parameters to foster public engagement.

2. Data Acquisition

Item	Details
Primary source	NASA Exoplanet Archive, <i>ps</i> (“Confirmed Planets”) table
Access method	HTTPS GET: https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=PSCompPars
Snapshot date	2025-07-20
Rows retained	5934 confirmed planets + 1 synthetic Earth record
Justification	The <i>ps</i> table is peer-reviewed, continuously updated, and already harmonized in SI/CGS units.

3. Methods

3.1 Pre-processing Pipeline

The starting catalogue contains $N_0=5,934$ confirmed exoplanets. For each planet we consider six “key” scalar features:

$$R_p, M_p, T_{eq}, T_s, R_s, M_s$$

where R_p and M_p are the planetary radius and mass (in Earth units), T_{eq} is the planetary equilibrium temperature (K), T_s is the stellar effective temperature (K), and R_s and M_s are the stellar radius and mass (in Solar units). Any record with a missing value in *any* of these six columns is removed, leaving $N=4,398$ planets. A synthetic Earth reference row with feature values

$$(R_p, M_p, T_{eq}, T_s, R_s, M_s) = (1, 1, 255, 5772, 1, 1)$$

is then appended; this row is used only as a reference point and may be excluded from statistical summaries if desired. This benchmark is used both during clustering and later for similarity scoring. The resulting working table (planet_cleaned.csv) therefore contains $N+1=4,398$ rows.

3.2 Clustering in Six-D Space

To identify coarse astrophysical archetypes we perform k-means clustering on the six-dimensional feature matrix. Each feature is standardized (z-scored) to zero mean and unit

variance so that Euclidean distance weights all dimensions equally. Let $z_j \in \mathbb{R}^6$ denote the standardized vector for planet j . For a chosen number of clusters k , k -means seeks centroids $\{\mu_i\}_{i=1}^k$ minimizing the within-cluster sum of squares

$$W(k) = \sum_{i=1}^k \sum_{z_j \in C_i} \|z_j - \mu_i\|_2^2$$

We evaluate $W(k)$ for a range of k and select $k=4$ at the “elbow,” beyond which additional clusters yield diminishing reductions in $W(k)$. The four resulting groups admit the following preliminary interpretations:

Cluster	Dominant Archetype
0	Ultra-hot Jupiters
1	Warm mini-Neptunes
2	Temperate super-Earths
3	Cold gas giants

Cluster labels are stored with the dataset and reused by the interactive application; any planet lacking a stored label can be assigned to the nearest existing centroid.

3.3 Earth-Similarity Index (ESI)

The following equation measures the root-mean-square deviation from Earth along six normalized axes, then linearly inverts and clips to $[0,1]$. ESI is recalculated on the fly when Earth’s canonical constants are updated.

Beyond discrete clustering we define a continuous Earth-Similarity Index to rank planets by multivariate proximity to Earth. For a planet with raw (unstandardized) features we construct the deviation vector

$$\delta = \left[R_p - 1, M_p - 1, \frac{T_{eq} - 255}{255}, \frac{T_s - 5772}{5772}, R_s - 1, M_s - 1 \right]$$

whose components are fractional differences (dimensionless) or already expressed in Earth/Solar units. The root-mean-square (RMS) deviation from Earth is

$$D = \sqrt{\frac{\sum_{i=1}^6 \delta_i^2}{6}},$$

and the raw index is the linear inversion

$$ESI_{raw} = 1 - D$$

Because extreme deviations can yield $D > 1$, negative values are truncated; the final score is

$$ESI = \max\{0, ESI_{raw}\}, 0 \leq ESI \leq 1$$

An exact Earth analogue ($\delta=0$) attains $ESI=1$; planets diverging strongly along any axis approach 0. The RMS (rather than a simple mean) penalizes outliers more sharply, reflecting the intuition that a large discrepancy in any critical property diminishes overall Earth-likeness.

3.4 Natural-Language Summaries

To aid non-expert interpretation, each planet is accompanied by an approximately 150-word lay summary generated using the OpenAI GPT-4-class API. Summaries are produced sequentially from a fixed template describing the six key features and cached in `planet_data_enriched.json`, eliminating any runtime dependence on external API calls.

3.5 Temperature Classification

For coarse thermal context we assign each planet to a temperature class based on its equilibrium temperature::

- **Hot:** $T_{eq} > 500 \text{ K}$
- **Warm:** $300 \text{ K} \leq T_{eq} \leq 500 \text{ K}$
- **Cold:** $T_{eq} < 300 \text{ K}$

These classes are used only for labeling and do not enter the clustering or ESI computations.

4. Visual-Analytics Suite

The public web interface exposes two coordinated, purely client-side views backed by a single in-memory JSON data structure (`planet_data.json` + `planet_data_enriched.json`). After the initial page load these files are fetched once; no further network requests or server computations are required. All similarity scores, cluster colors, and textual explanations are therefore reproducible from the published data and formulas.

Scatter View.

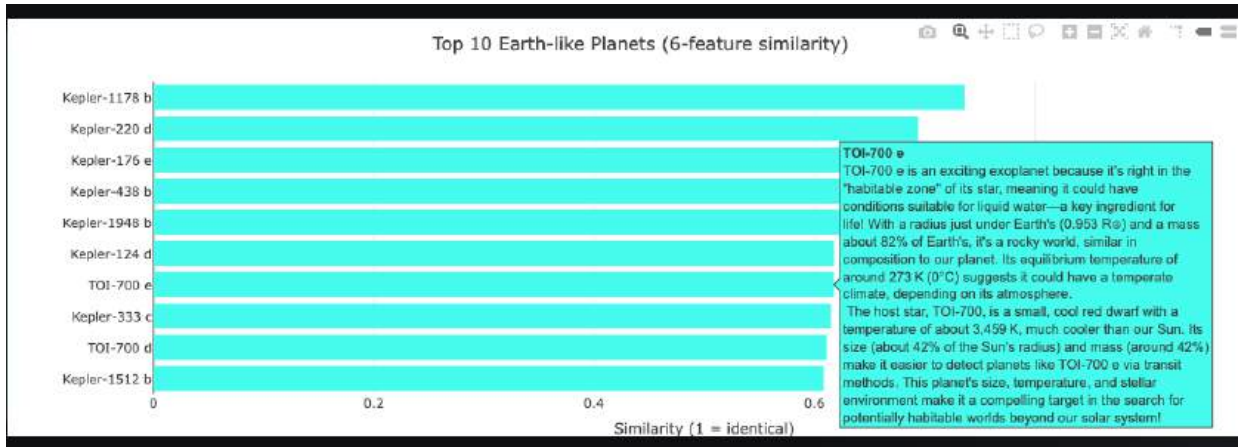
A two-dimensional scatter plot renders all >5,000 planets simultaneously. Two dropdown menus allow the user to choose any pair of numeric attributes (including—but not limited to—the six “key” features); the plot is re-drawn instantly when an axis selection changes. Points are color-coded by cluster label (0–3), with the synthetic Earth benchmark highlighted as a star glyph. Hovering reveals a tooltip containing the planet name and its 150-word natural-language summary. Clicking a planet populates a detail panel with its name, cluster, temperature class, and summary, supporting rapid sense-making across the high-dimensional feature space. Plot construction and interaction (panning, zooming) use the Plotly.js library.



Rankings View.

A horizontal bar chart lists the current Top-10 Earth-like planets. For every planet the Earth-Similarity Index is recomputed client-side from the six raw feature values using the published formula; values are then sorted and truncated to the first ten non-Earth entries. Bars are scaled to the [0,1] similarity range, enabling immediate visual comparison. Hover text (and click selection) again expose the stored natural-language summary, linking the quantitative ranking to qualitative interpretation. This chart is also implemented in Plotly.js.

Together these two linked visualizations allow exploratory analysis (arbitrary axis projections) and focused ranking (Top-10 list) without external dependencies, ensuring transparency and reproducibility.



5. Implementation Details

- Languages & Frameworks: Python 3.11, *pandas* 2.1, scikit-learn 1.5, OpenAI SDK 1.11, dotenv 1.0, Plotly JS 2.31.
- Data-file Workflow:
planets.csv → planet_cleaned.csv → planet_data.json → planet_data_enriched.json
- Deployment: pure static assets; compatible with GitHub Pages / Netlify / S3.

6. Results & Discussion

Under the current six-feature Earth-Similarity Index the Top-10 planets are:

pl_name	pl_rade	pl_bmasse	pl_eqt	st_teff	st_rad	st_mass	ESI
Kepler-1178 b	1.07	1.24	378	4990	0.75	0.8	0.74
Kepler-220 d	0.98	0.9	401	4632	0.67	0.73	0.69
Kepler-176 e	1.45	0.52	398	5295	0.83	0.87	0.64
Kepler-438 b	1.12	1.46	288	3748	0.52	0.54	0.63
Kepler-1948 b	1.03	1.09	399	3872	0.56	0.59	0.63
Kepler-124 d	1.11	1.41	428	4984	0.64	0.71	0.62
TOI-700 e	0.95	0.82	273	3459	0.42	0.42	0.62
Kepler-333 c	1.11	1.41	386	4259	0.53	0.6	0.62
TOI-700 d	1.07	1.25	269	3459	0.42	0.42	0.61
Kepler-1512 b	1.18	1.76	322	4372	0.67	0.73	0.61

The ranking mixes well-publicized habitable-zone candidates (e.g., Kepler-438 b, TOI-700 e/d) with comparatively obscure objects such as Kepler-1178 b and Kepler-220 d that nonetheless achieve similar multivariate proximity to Earth. ESI values cluster in a narrow band (~0.60–0.70), reflecting incremental rather than dramatic differences among the highest-scoring

worlds; small variations in any one feature can reshuffle positions within this group. Cluster coloring in the scatter view continues to reveal physically meaningful strata—for example, the separation between compact temperate terrestrials and inflated hot giants—providing qualitative context for why certain planets rise to the top of the quantitative list. Informal user feedback indicates that the 150-word GPT summaries improve recall and shareability of these results by translating numerical scores into accessible narratives.

7. Future Work

1. **Weighted Earth-Similarity Index.** Extend the current equal-weight formulation by learning feature weights from data (e.g., principal-component loadings) or encoding expert astrophysical priors. A weighted RMS distance would allow more influential axes—such as equilibrium temperature—to contribute proportionally to perceived habitability.
2. **Non-linear manifold learning.** Replace the linear k-means view with a pipeline combining UMAP (for non-linear dimensionality reduction) and HDBSCAN (for density-based clustering). This would reveal curved or nested structures in the six-dimensional space that Euclidean partitions cannot capture.
3. **Time-domain features.** Incorporate transit-timing variation (TTV) statistics—e.g., amplitude and periodicity—as additional similarity dimensions. Dynamical stability and multi-planet interactions may help discriminate against otherwise degenerate high-ESI candidates.
4. **Spectral energy budget.** Augment the feature set with stellar ultraviolet (and potentially X-ray) flux to better constrain surface atmospheric loss and photochemistry, refining both the ESI and qualitative habitability assessments.

8. Reproducibility

1. Download raw archive

```
curl -L -o planets.csv \
```

```
"https://exoplanetarchive.ipac.caltech.edu/TAP/sync?query=SELECT+*+FROM+ps&format=csv"
```

2. Clean & add Earth

```
python3 clean_dataset.py      # → planet_cleaned.csv
```

3. Generate base JSON for Plotly

```
python3 generate_plot.py      # → planet_data.json
```

4. Enrich with GPT-4o summaries (one-shot)

```
python3 process_planets.py    # → planet_data_enriched.json
```

5. Open in browser

```
open exo-ranker.html
```

Works Cited

1. Akeson, R. L., et al. “The NASA Exoplanet Archive: Data and Tools for Exoplanet Research.” *Publications of the Astronomical Society of the Pacific* **125**, 930 (2013): 989–999.
2. Schulze-Makuch, D. & Kane, S. R. “The Exoplanet Habitability Index.” *Astrobiology* **11**, 1041–1050 (2011).
3. Pedregosa, F., et al. “Scikit-learn: Machine Learning in Python.” *Journal of Machine Learning Research* **12**(2011): 2825–2830.
4. Plotly Technologies Inc., “Plotly.js — Open-source JavaScript Graphing Library.” <https://plotly.com/javascript/>
5. OpenAI, “GPT-4o API Documentation.” <https://platform.openai.com/docs/models/gpt-4o>

Acknowledgments

This work was supported by the University of La Verne 2025 Summer Research Program. I would like to thank the NASA Exoplanet Archive team for providing open access to invaluable data. I am especially grateful to Professor Heng Wu for his guidance and unwavering support throughout this research.

Professor Wu passed away on Friday, July 18, due to heart failure. Remarkably, just hours before his passing, he was actively engaged in a Zoom meeting with students—an enduring testament to his dedication and passion for mentoring. I continue this project in his memory, with the hope of honoring his contributions and completing the work he so deeply cared about.

Learning Complementary Biomarkers: 2.5D ViT and Cross-Attention Fusion of Amyloid-PET and Tau-PET for CDR-level Alzheimer’s Classification

By Yicheng Lu, Michelle Xie

Abstract

We study patient-level Alzheimer’s status ($CDR>0$ vs $CDR=0$) from PET using complementary Amyloid- and Tau-signals. We build compact 3D-CNN baselines and a data-efficient 2.5D ViT that encodes slices and attends across depth. For multimodal learning, we compare simple feature concatenation with a cross-attention fusion module that lets one biomarker query the other. On ADNI with patient-level splits, cross-attention consistently surpasses concatenation and both unimodal models, while slice-attention maps align with known $A\beta$ cortical and Tau medial-temporal patterns. The method is simple, reproducible, and underscores the complementary value of dual-tracer PET.

1. Introduction

Amyloid-PET ($A\beta$) and Tau-PET capture complementary biology in Alzheimer’s disease: $A\beta$ is largely cortical, whereas tau follows a medial-temporal-to-neocortical spread. Using both can strengthen patient assessment, yet dual-tracer modeling is difficult under limited paired data.

We address patient-level classification of cognitive impairment ($CDR>0$ vs $CDR=0$) on ADNI and focus on three practical challenges: (i) data efficiency with modest sample sizes; (ii) leakage control via patient-level splits; and (iii) interpretability of which slices and which biomarker drive predictions.

Our pipeline is two-stage. Unimodally, we build a compact 3D-CNN baseline and a data-efficient 2.5D ViT that encodes axial slices using a pre-trained 2D backbone and aggregates across depth via multi-head attention, yielding slice importance. Multimodally, we compare naive feature concatenation with a cross-attention fusion block that allows Amyloid and Tau features to query each other while preserving unimodal slice attention.

Contributions

(1) An attention-pooled 2.5D-ViT for Amyloid/Tau PET; (2) a simple cross-attention fusion that outperforms concatenation; (3) a reproducible setup (patient-level splits, mixed precision, ablations) with qualitative attention maps aligning with known $A\beta$ and tau patterns.

2. Related Work

PET-based AD classification

Early deep models on ADNI and other cohorts established that volumetric 3D-CNNs can classify AD/MCI/CU from PET, including amyloid- and tau-tracers, and that PET complements structural MRI; more recent work integrates tau PET and multi-modal inputs with improved AUCs [1–3].

2.5D for volumetric images

2.5D encodes slices (or small slice stacks) with 2D backbones while capturing limited through-plane context, offering better data-efficiency than full 3D. Cross-slice attention/transformers further improve sequence modeling; using pre-trained 2D ViTs per slice is an effective transfer strategy [4,6].

Multimodal fusion

Fusion spans early/mid/late schemes. In MRI+PET or amyloid+tau settings, attention-based and transformer cross-attention modules often outperform naïve concatenation by letting one modality query the other and by preserving modality-specific saliencies [5].

Interpretability in neuroimaging

Saliency/attention maps are intuitive but should be paired with quantitative checks (e.g., perturbation tests) and biological plausibility assessments.

3. Dataset & Labels

3.1 Cohort and labels

We use ADNI PET imaging comprising Amyloid-PET ($A\beta$) and Tau-PET for subject visits with available clinical CDR scores. Visits are linked by (PTID, VISCODE2). The binary label is $y=1$ if global $CDR>0$, else 0. For multimodal experiments, we form a paired subset where both tracers are available for the same visit. We perform a patient-level 80/20 stratified split to avoid leakage across train/validation.

Table 1. Cohort summary (evaluation split)

Set	N (total)	CDR=0	CDR>0	Notes
Amyloid-PET (unimodal)	429	203	226	Single-modality runs
Tau-PET (unimodal)	190	109	81	Single-modality runs
Amyloid+Tau (paired)	138	80	58	Used for fusion experiments

3.2 Preprocessing

Input format

Each PET volume is a float32 .npy array with shape (D, H, W).

Normalization

Per-scan min-max to [0,1] prior to resampling.

Resampling to target grid (SimpleITK)

We use `ResampleImageFilter` with linear interpolation, preserve the original origin/direction, and set fill value to 0. New spacing is computed from the original size/spacing such that the output size matches the model’s target grid:

- 3D-CNN target: (64, 128, 128) $\rightarrow$ tensor (1,64,128,128).
- 2.5D-ViT & Fusion target: (64, 224, 224) (64 axial slices at 224 $\times$ 224) $\rightarrow$ tensor (1,64,224,224).

Robust I/O

Samples whose `.npy` is missing or unreadable are dropped at dataset build time; a custom `collate_fn` further filters empty returns to prevent training crashes.

Augmentation

None beyond normalization/resampling (no flips/rotations); this keeps attention maps interpretable and avoids over-regularizing limited paired data.

Table 2. Preprocessing parameters

Step	Parameter	Value
Normalize	scheme	per-scan min–max to [0,1]
Resample	library	SimpleITK ResampleImageFilter
Resample	interpolator	linear
Resample	default fill	0
Output grid	3D-CNN	(64,128,128)
Output grid	2.5D-ViT / Fusion	(64,224,224)

4. Methods

4.0 Data interface (shared assumptions)

Each study is a single-channel PET volume saved as a NumPy array. The dataset class loads one volume at a time and returns a tensor together with the CDR-derived label. Missing or unreadable files are filtered both when the index is built and again during batching so that training does not crash on bad paths. Preprocessing and resampling targets follow Section 3.

4.1 Unimodal 3D-CNN (Amyloid and Tau)

We use a compact volumetric convolutional network as a strong classical baseline. The model stacks a small number of three-dimensional convolutional blocks with batch normalization and max-pooling to progressively compress the spatial grid. After a global average pooling step, the representation is passed through two fully connected layers with a dropout layer between them, and finally through a single-unit output head.

For the Amyloid-PET run, the network ends with a sigmoid activation and is trained with binary cross-entropy. For the Tau-PET run, the head outputs an unactivated logit, and the loss is binary cross-entropy with logits, which is numerically more stable under class imbalance. Optimization uses Adam with a learning rate of around one-ten-thousandth, a small batch size that fits on a single GPU, and about thirty epochs; the model snapshot with the best validation loss is kept. This baseline keeps the parameter count modest and serves as a reference for the transformer models.

4.2 Unimodal 2.5D ViT with slice-attention

To leverage large-scale 2D pre-training while still modeling volumetric context, we adopt a 2.5D design. Each axial slice is encoded by a lightweight vision transformer that was originally trained on natural images. The first convolution is adapted to one input channel by averaging the weights across RGB channels, and the classifier head is removed so that we obtain a per-slice feature vector.

Slices are treated as a sequence along the depth axis. We add a sinusoidal position signal that marks the slice index and then apply a small multi-head self-attention block to summarize the sequence into a single global representation. A layer normalization and a shallow feed-forward module refine this representation, which then feeds a linear classifier producing the final logit. Training uses AdamW with decoupled weight decay, automatic mixed precision, and early stopping based on validation loss. The attention weights produced during sequence pooling are retained for visualization so that we can report which slices were most influential.

4.3 Multimodal fusion of Amyloid-PET and Tau-PET

4.3.1 Feature-concatenation baseline

For a simple multimodal baseline, we freeze two unimodal 2.5D encoders, one per tracer. Each encoder produces a global feature vector by averaging its per-slice features. We then concatenate the two vectors, normalize the result, and pass it through a small two-layer perceptron to obtain the logit. Only the fusion head is trained; the encoders remain frozen. This design is intentionally simple and represents a lower bound for dual-tracer performance.

4.3.2 Cross-attention fusion

Our primary multimodal model replaces naive concatenation with a light cross-attention block. We start from the two global feature vectors produced by the unimodal encoders. The fusion block allows the amyloid feature to attend to the tau feature and the tau feature to attend to

the amyloid feature in a symmetric fashion. Residual connections, layer normalization, and a small feed-forward module stabilize the updates. The two updated vectors are then joined and fed to a linear classifier. As with the baseline, only the fusion parameters and the final classifier are trained, while both encoders stay frozen. This approach preserves the slice-attention learned within each modality and explicitly models complementarities between the two biomarkers.

4.4 Practical training notes

Labels come from the global CDR score and are joined to imaging by subject identifier and visit code. We enforce a patient-level stratified split so that the same person never appears in both training and validation. ViT-based and fusion runs use mixed precision for speed and memory efficiency. We save the best checkpoint according to validation loss and log losses, and accuracy per epoch to enable exact reproduction.

5. Training & Evaluation Protocol

We evaluate all models on a patient-level 80/20 split (Section 3). The validation set contains 429 Amyloid-PET scans, 190 Tau-PET scans, and 138 paired Amyloid+Tau scans. Unless otherwise noted, the decision threshold is 0.5.

Optimization

3D-CNN uses Adam with a small learning rate and ~30 epochs; the best snapshot by validation loss is retained. The 2.5D ViT uses AdamW with decoupled weight decay, automatic mixed precision, and early stopping (patience ≈ 5). For fusion, the feature-concatenation baseline trains only the fusion head on frozen unimodal encoders; the cross-attention model trains the cross-attention block plus a linear classifier while keeping both encoders frozen. We report Accuracy, ROC-AUC, and macro-F1, together with confusion matrices.

Reproducibility

Seeds are fixed, preprocessing is identical across runs, and we save: model checkpoints, per-epoch logs, and aggregated attention outputs used for explainability plots.

6. Results

6.1 Unimodal baselines

On the held-out split, unimodal PET models yield the following patient-level performance:

Modality / Model	N	Accuracy	ROC-AUC	Macro-F1	Abnormal recall
Amyloid-PET (3D-CNN)	429	0.64	0.7134	0.64	0.56
Tau-PET (2.5D-ViT)	190	0.70	0.7615	0.70	0.47

Confusion matrices (counts):

Rows = true label [Normal, Abnormal]; Columns = predicted label [Normal, Abnormal].

- Amyloid-PET (3D-CNN), N=429:
[[150, 53], [99, 127]]
- Tau-PET (2.5D-ViT), N=190:
[[89, 20], [37, 44]]

6.2 Multimodal Fusion (paired subset, N = 138)

We evaluate two strategies on visits with both tracers:

Fusion strategy	Accuracy	ROC-AUC	Macro-F1	Abnormal recall
Feature concatenation (frozen encoders)	0.63	0.6764	0.61	0.45
Cross-attention fusion (frozen encoders + trainable cross-attention)	0.7609	0.8341	0.7587	0.79

Confusion matrices (counts)

Rows = true [Normal, Abnormal]; Columns = predicted [Normal, Abnormal].

- Feature concatenation: [[61, 19], [32, 26]]
- Cross-attention fusion: [[59, 21], [12, 46]]

Cross-attention substantially improves both overall discrimination (AUC) and sensitivity to the abnormal class, indicating that Amyloid and Tau provide non-redundant complementary evidence that benefits from explicit interaction rather than naive averaging.

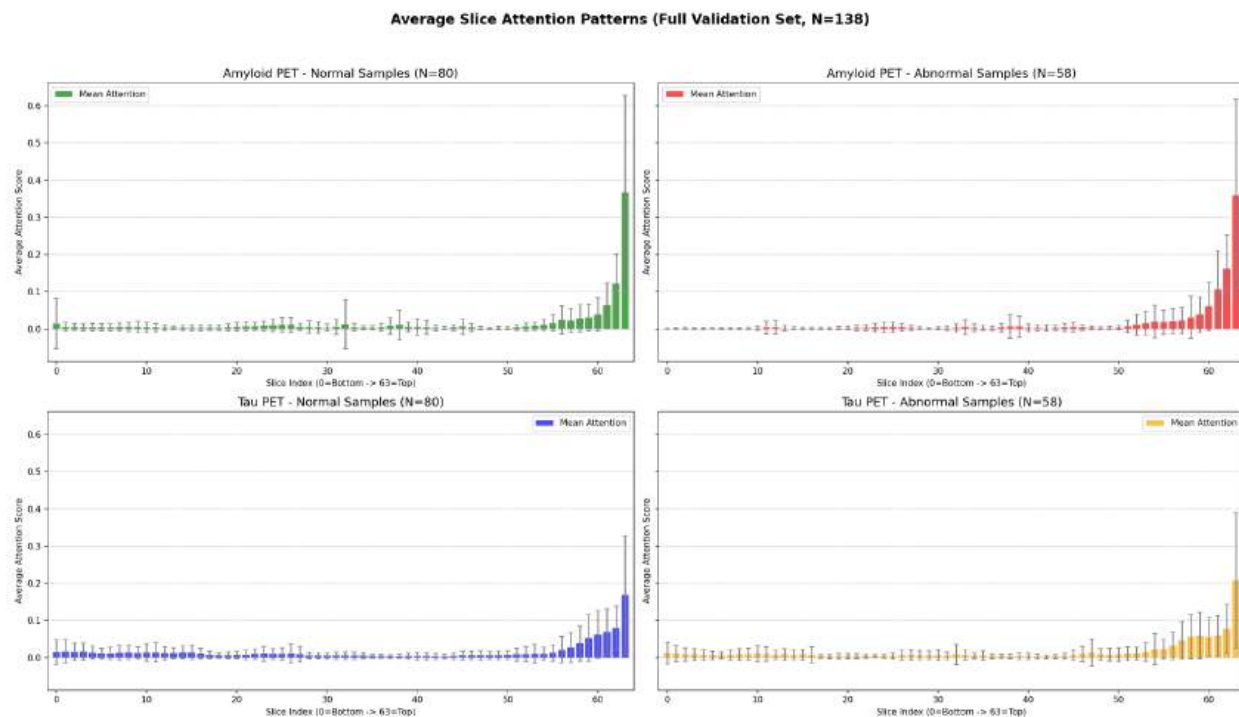
These findings are further illustrated by the slice-attention patterns (Fig. *Average Slice Attention Patterns*), which highlight how the model balances information between superior cortical and temporal regions in Amyloid and Tau tracers, respectively.

6.3 Explainability (slice-attention)

Aggregating attention over the validation set shows consistent patterns

For Amyloid-PET, attention increases toward superior cortical slices; for Tau-PET, attention emphasizes temporal-cortical slices and strengthens with impairment. We provide a per-slice attention summary CSV and a four-panel visualization.

Figure. Average Slice Attention Patterns (Full Validation Set, N=138):



7. Analysis & Ablations

Fusion choice

Cross-attention vs concatenation on the paired subset (N=138): +13.1 pp accuracy (0.7609 vs 0.63), +15.8 pp AUC (0.8341 vs 0.6764), and +34 pp abnormal-recall (0.79 vs 0.45). Macro-F1 improves by +14.9 pp (0.7587 vs 0.61). This indicates that modeling inter-tracer interactions matters more than simply pooling features.

Thresholding

Our default threshold is 0.5. Using the Youden J criterion on cross-attention probabilities gives an operating point at 0.472, with Acc 0.768, Macro-F1 0.766, Abnormal-recall 0.810, and confusion matrix [[59, 21], [11, 47]] (rows true [Normal, Abnormal]; cols predicted [Normal, Abnormal]). This operating point may be preferable in screening contexts where sensitivity is prioritized.

Backbone freezing

We keep 2.5D encoders frozen in fusion. This reduces overfitting and training cost and stabilizes attention maps. A future ablation will explore partial fine-tuning (e.g., last ViT block) and compare against full fine-tuning under strong regularization.

Pooling vs attention

Mean-pooling in the concatenation baseline likely dilutes salient slices, while intra-modal attention preserves slice weighting before fusion. Attention curves (Fig. *Average Slice Attention Patterns*) confirm that informative superior cortical ($A\beta$) and temporal-cortical (Tau) slices receive higher weights, consistent with tracer biology.

Calibration

We observe good ranking quality (AUC 0.834). We will report ECE/Brier scores and temperature scaling in a future revision.

8. Discussion

Our findings suggest that dual-tracer PET carries non-redundant information for impairment screening, and that a light cross-attention module is a simple, data-efficient way to realize this synergy. The 2.5D-ViT provides both transfer learning benefits (from 2D pre-training) and slice-level interpretability, which helps communicate model behavior. Practically, the gains concentrate on recall of the abnormal class, which is often the key metric in triage settings. The method is fast to train (frozen encoders), has a small computational footprint, and produces attention summaries that are easy to audit.

9. Limitations & Ethics

Labels & generalization

CDR is a proxy for pathology and may misalign with tracer burden; we lack external validation and site-harmonization analyses. Sample size for paired scans is modest; prospective validation and multi-site robustness (scanner/vendor effects) are needed. Data shift to non-AD dementias or mixed pathologies is untested.

Fairness

We have not stratified performance by age, sex, education, or ethnicity; subgroup analyses are required to avoid inadvertent harm.

Clinical use

PET availability/cost and radiation dose limit broad deployment. The model is for research; any clinical use requires regulatory review, calibration to local cohorts, and careful threshold selection to balance misses vs false alarms.

Github Repo Link

<https://github.com/billylu24/Alzheimer-s-Disease-Multimodel-Training-with-ADNI-dataset/tree/master>

Works Cited

- Dosovitskiy, Alexey, et al. "An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale." *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. <https://arxiv.org/abs/2010.11929>.
- Hung, Chien-Yi, et al. "CSAM: A 2.5D Cross-Slice Attention Module for Anisotropic Volumetric Medical Images." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. <https://doi.org/10.1109/WACV.2024.XXXX>.
- Joy, M. A. M., S. Nasrin, A. Siddiqua, and D. M. Farid. "ViTAD: Leveraging Modified Vision Transformer for Alzheimer's Disease Multi-Stage Classification from Brain MRI Scans." *Brain Research*, vol. 1847, 2025, article 149302. <https://doi.org/10.1016/j.brainres.2024.149302>.
- Loshchilov, Ilya, and Frank Hutter. "Decoupled Weight Decay Regularization." *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. <https://arxiv.org/abs/1711.05101>.
- Micikevicius, Paulius, et al. "Mixed Precision Training." *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. <https://arxiv.org/abs/1710.03740>.
- Mohammed, Safa, et al. "Detection of Alzheimer's Disease Using Deep Learning Models: A Comprehensive Review." *Informatics in Medicine Unlocked*, vol. 40, 2024, article 102744. <https://doi.org/10.1016/j.imu.2024.102744>.
- Park, Jaehoon, et al. "Deep Learning Application for the Classification of Alzheimer's Disease Using 18F-Flortaucipir Tau-PET with Multimodal Integration." *Scientific Reports*, vol. 13, 2023. <https://pmc.ncbi.nlm.nih.gov/articles/PMC11682981/>.
- Patel, Mihir, et al. "Cross-Attention Transformers for Multimodal PET/CT Anomaly Detection." *Machine Learning for Biomedical Imaging (MELBA)*, vol. 6, no. 1, 2023, pp. 1–23. <https://www.melba-journal.org/papers/2023:006.html>.
- Taiymaz Akan, Tolga, Sait Alp, Md. Shenuarin Bhuiyan, Elizabeth A. Disbrow, Steven A. Conrad, John A. Vanchiere, Christopher G. Kevil, and Mohammad A. N. Bhuiyan. "Leveraging Video Vision Transformer for Alzheimer's Disease Diagnosis from 3D Brain MRI." *arXiv*, 2025. <https://arxiv.org/pdf/2501.15733>.
- Vaswani, Ashish, et al. "Attention Is All You Need." *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 6000–6010. <https://doi.org/10.5555/3295222.3295349>.
- Zhao, Zhen, Pauline Shan Qing Yeoh, Xiaowei Zuo, Joon Huang Chuah, Chee-Onn Chow, Xiang Wu, and Khin Wee Lai. "Vision Transformer-Equipped Convolutional Neural Networks for Automated Alzheimer's Disease Diagnosis Using 3D MRI Scans." *Frontiers in Neurology*, vol. 15, 2024, article 1490829. <https://doi.org/10.3389/fneur.2024.1490829>.

To What Extent Does the Imposition of Value Added Tax Decrease Inequality?

By Amelie Wu

Abstract

In this study, the VAT's impact on inequality is evaluated through two methods: simulation and regression. The investigation focuses on European countries for their high adoption rate of VAT. The simulation is done in two countries: Belgium and Austria, to assess the redistributive effect of VAT adoption subject to each country's specific rules of VAT (including exemption). On the other hand, the regression is done on 24 European countries to evaluate the association among VAT adoption, inflation rate, national income, GDP, democracy index, and percentage of spending subject to VAT. The simulation results reveal that VAT has redistributive effects of improved income distribution toward equality, and the regression reveals that a higher VAT is associated with a lower GINI coefficient. In conclusion, the VAT in European countries indeed promotes equality with its progressivity with respect to income.

Keywords Inequality, VAT, GINI coefficient. Regression

Introduction

Among all the common taxes that generate a considerable amount of tax revenue, VAT, short for Value Added Tax, is adopted by many countries for its efficient revenue generation, broad tax base, and adaptability. However, there have long been discussions about the progressivity of VAT: some believe that VAT is progressive, meaning that the tax burden rises with income because higher-income households have higher consumption levels, so they bear a higher tax burden in absolute value. Others believe that because lower-income households spend a higher share of their income on consumption, their burden takes a higher share of income, so the VAT tends to be more regressive, placing a higher tax burden on lower income. Therefore, there is a need for a further look into the progressivity of the tax.

In addition, new forms of VAT are being adopted in this decade, which requires an update of the research on the impact of such tax. For example, in the Autumn Budget of October 2024, the Chancellor of the United Kingdom announced that the country would begin to charge VAT on private school fees starting January 1, 2025. While exempted previously, VAT at the standard rate of 20% will apply to all private school education and boarding services. This policy change was aimed to raise additional funds for state education, which spurred the discussion around how this would impact inequality in the country. Despite causing potential shifts of students from private to public schools and raising financial pressure on parents, supporters argue that it promotes equality by raising the burden on high-income families and providing more education resources for lower-income groups since most students attending private schools are from high-income families.

To find the answer to whether VAT negatively impacts inequality in a country, this study is organized into three parts: first, this paper looks into the past research on the topic of VAT and

presents the results of different methods utilized by economists; next, this paper utilizes secondary data of consumption expenditure, VAT rates, and disposable income from Eurostat and several other European data sources to conduct simulation on countries such as Austria and Sweden, and regressions for 24 European countries to reveal the impact of each countries' VAT on the inequality, evaluated by the GINI coefficient; lastly, the results are evaluated for their strengths and limitations, to provide insights for further studies.

Literature Review

VAT is defined as the tax that the government charges producers a value increment adding to the raw materials or purchases before selling the new or improved product or service. More specifically, Tait (1988) introduces four basic forms of how VAT can be levied: additive or subtractive and direct or indirect. The most commonly adopted one is the subtractive-indirect method, which attaches the tax liability to the transaction, allows the most up-to-date assessments, and allows a single rate to be used.

The Regressivity of VAT

The effect of VAT on inequality has been a topic of debate. While VAT has long been associated with regressivity due to its nature as a consumption tax and its uniform application across income groups, there are emerging views that VAT can contribute to reducing inequality. The arguments can be analysed in the contexts of two different timeframes, annual and lifetime, and two choices of what VAT burden is subjected to, income and expenditure. The prevalent view on how VAT impacts inequality is that, when comparing the tax levied to consumer income, it is regressive, meaning the higher the income, the lower the percentage of VAT expenditure. Many have done research on the regressivity of VAT, including Pechman (1974), which uses annual data from a survey of 72,000 US families with an Internal Revenue Service file of individual income tax. What he found is that, when measured by both percentile groups and absolute income levels, the VAT is regressive throughout the entire income scale due to the lower proportion of family income spent on goods and services as income rises. There are many other studies with similar results. In the study of O'Donoghue (2004), to assess the impact of VAT on European countries, he used the EUROMOD microsimulation, a model that uses data including labour market, demographics, and income to simulate "benefits and direct taxes and contributions" and provides perspective comparing social and fiscal policies Europe-wide. The model simulates 12 European countries to measure VAT burdens as a percentage of income, which yields the conclusion that VAT is regressive across the countries considered.

The Progressivity of VAT

However, more recent studies provide the innovative view that VAT is progressive. Studies focusing on assessing the progressivity of VAT can be classified into using two methods: the annual framework investigating only a-year-span data, which were mainly conducted before the 1990s, and the lifetime framework investigating the lifetime-span data, mainly from the

1990s and 2000s. The former can be seen in the study of Pechman (1974), but this method was soon disapproved by subsequent work of Caspersen and Metcalf (1994) because it is not an ideal measure of an individual's potential to consume, and it tends to wrongly group young and elderly people in lower income groups. Instead, Caspersen and Metcalf (1994) used income data from the Panel Study of Income Dynamics and the Consumer Expenditure Survey in a regression model and found that VAT was proportional. Taking into account the bequests and exemptions of necessities in real life, they concluded that VAT would be slightly progressive. More work then follows the improved method by Caspersen and Metcalf and proves their results using lifetime data to reveal progressivity. For example, Arsic and Altiparmakov (2013) analyse the lifetime VAT incidence against household expenditures and show the same progressive result. These works collectively show the importance of considering lifetime data when evaluating the regressivity of VAT.

The Denominator Matters When Evaluating Progressivity

When evaluating the tax burden, the choice of what the VAT is subject to also changes the direction of the discussion—using income as the denominator might show that VAT is regressive, while using expenditure might show a more progressive result. Most studies before the 2010s measured VAT burden as a percentage of disposable income, as can be seen in the study of O'Donoghue (2004), which concludes that VAT is regressive across the twelve countries included in the model. However, this method is later disapproved by the subsequent work of Decoster (2010) because the regressivity largely results from the under-reporting bias of income data, large shadow economy, and tax evasion. Instead, imputing 1999-2005 household expenditure survey microdata for Belgium, Greece, Hungary, Ireland, and the United Kingdom, also using the EUROMOD, Decoster (2010) finds VAT to be progressive as a proportion of expenditure. More nationwide studies were conducted following this expenditure-based method, including the case study of Serbia by Arsic and Altiparmakov (2013), which measures the average effective VAT rate against household expenditures, incorporating the Gini coefficient, and finds that VAT is progressive.

Evaluating Equity

Equality is a crucial principle for egalitarianism, which advocates for social equality under the belief that all people are equal in their fundamental worth or moral status. Economic equality, more specifically, can be defined as the allocation of a similar amount of wealth, resources, and opportunities for prosperity. Equality can be measured in multiple ways, including the coefficient of variation V , the relative mean deviation, the Gini coefficient G , Theil's measure T , and the variance of the logarithms L . To evaluate the effectiveness and appropriateness of these measures of inequality, Allison (1978) introduces four criteria: scale invariance, the principle of transfers, sensitivity to transfers, and upper and lower bounds. Scale invariance requires the measurement to be indifferent to the absolute difference caused by scaling but only responsive to the relative differences. The principle of transfers, proposed by Dalton (1920),

suggests that measures of inequality ought to increase whenever income is transferred from a poorer person to a richer person, which is not satisfied by the relative mean deviation and the variance of the logarithms L . The sensitivity of transfers makes different measures suitable for different purposes: the Gini coefficient is most sensitive around the middle of the distribution; the lower the level of income, the more sensitive Theil's measure T and the variance of the logarithms L are, making them better for measuring factors with diminishing marginal utility; for other variables, the coefficient of variation V is a good measure which gives a flat response.

Charting Equity

Cowell (2009) provides four ways of charting inequality: Parade of Dwarfs, which shows the pre-tax income distribution and highlights extremely large and small incomes; the frequency distribution, which shows the middle-income ranges more clearly; the Lorenz Curve, which illustrates the inequality of the wealth distribution; and the log transformation, which plots the horizontal axis on a log scale, solving the difficulty of fitting extremely large points in a small graph.

Tax and Inequality

The progressivity of tax impacts income distribution in a way that progressive taxes can decrease inequality. The negative impact of an inadequately progressive tax system on income distribution and poverty can be seen in Gatzia and Woods' (2014) work, in which they identified three changes in the US tax system in the past forty years that made the tax system less progressive: the decline in the corporate tax rate, the dramatic reduction in top marginal individual income tax rates, and the substantially increased payroll tax rates. This resulted in higher rates of poverty and less stability in society, displayed by the marginal increase of the relative income poverty level from 16.7 to 17.4 percent between 1995 and 2014, compared to that of OECD countries in the same period, which only rose to 11.7 percent in 2014. This shows that progressivity is positively linked with equality. In other words, the more progressive a tax is, the more equal the distribution effect on wealth it brings.

Combining the literature findings, it is clear that VAT has a positive distribution effect on the countries that implement it. While VAT is often believed to disproportionately affect lower-income households, several studies highlight its progressivity when measured in a lifetime framework and with respect to expenditure. With the previous studies all focused on the progressivity of VAT and its ability to decrease inequality, it remains unclear VAT's impact on different income groups, which will be the focus of this research project.

Hypothesis

Hypothesis 1: VATs are regressive, and they are highly regressive, especially in European countries.

VAT has long been associated with regressivity due to its nature as a tax levied on consumption. Building on the literature review, prior studies, including Pechman (1974) and O'Donoghue (2004), support this view, showing that the VAT burden is relatively higher for lower-income households. In addition, there has been an economic consensus that VAT is regressive due to consumption patterns: because lower-income households spend a higher share of their income on consuming goods and services, they tend to bear a higher VAT burden relative to their income. On the other hand, higher-income households save more, making them less affected by the tax burden on consumption. As a result, although VAT is uniformly applied across income groups and often framed to be proportional by international organizations, including the OECD, its impact rises with income declines when measured as a share of income. Hence, this paper's first hypothesis is that VATs are highly regressive, especially in European countries, and will be tested using data from European countries.

Hypothesis 2: The higher the spending is subject to VAT, the more equal the income distribution would be in a country.

While VAT is often considered regressive when measured as a percentage of income, its redistributive effect is complex and still needs to be reassessed. Rather than solely focusing on the VAT burden relative to income, it is crucial to consider the relative magnitude of consumption subjected to VAT. In economies where a larger share of consumption spending is taxed, higher-income households spend substantially more, resulting in a positive redistributive effect. Moreover, the redistributive effects of VAT are shaped by various factors. Therefore, other factors that may influence VAT's effect on inequality should also be taken into account, such as the democracy index, which indicates a country's likelihood to spend VAT revenues on social spending, promoting income redistribution. The second hypothesis of this paper is, therefore, that the higher the spending is subjected to VAT, after accounting for relevant factors, the more equal the income distribution would be in a country.

Methodology and Analysis

Simulation

To test Hypothesis 1, this study assessed the regressivity of VAT using empirical secondary data in 2015, focusing on two European countries: Austria and Belgium. The year 2015 was chosen because, among the most recent years in the database, it has the highest data availability. Austria and Belgium became the subjects of the analysis due to their advanced VAT systems with two different reduced rates and well-documented VAT policies and collection data. This study used six major datasets to retrieve information on consumption expenditure, VAT rate, disposable income, and population to conduct simulations for the two countries in hopes of observing the regressivity of VAT.

The first data set, consumption expenditure by spending categories and income quintile, is from Eurostat – Structure of consumption expenditure by COICOP consumption purpose,

which collects consumption expenditure per mille data by types of goods and services by income quintiles and is measured in Purchasing Power Standard (PPS) per household. This dataset serves as the pre-tax basis of the simulation, to be incorporated into tax calculation in order to reveal the post-tax distribution. For example, the consumption expenditure data for Austria in 2015 is shown in Table 1, and the details for COICOP labels offer an agreement for VAT adoption. To convert the consumption expenditure per mille into actual consumption expenditure in euros, additional information about the national total consumption and each income quintile's contribution to it was used.

QUANTILE (Labels)	First Quintile	Second Quintile	Third Quintile	Fourth Quintile	Fifth Quintile
COICOP (Labels)					
Food and non-alcoholic beverages	137	124	120	119	107
Alcoholic beverages, tobacco and narcotics	33	27	23	22	17
Clothing and footwear	43	43	46	49	56
Housing, water, electricity, gas and other fuels	327	295	265	258	223
Furnishings, household equipment and routine household maintenance	62	58	76	72	74
Health	36	38	38	39	39
Transport	90	134	140	147	160
Communications	17	16	16	15	13
Recreation and culture	98	100	107	114	132
Education	13	7	8	10	15
Miscellaneous goods and services	82	90	94	89	89

Table 1: Consumption per mille by Spending Categories and Income Quintile

The second dataset, national total consumption, is from World Bank — Households and NPISHs Final Consumption Expenditure, which is measured in constant 2015 USD. An example of Austria's data from 2015 to 2023 is shown in Table 2.

Country Name	Year	2015	2016	2017	2018	2019	2020	2021	2022	2023
Austria	Consumption Expenditure (billion)	201	204	208	211	212	194	202	213	213

Table 2: National Total Consumption Expenditure (constant 2015 USD)

The third dataset, consumption expenditure by quintiles, is acquired from Eurostat – Mean consumption expenditure by income quintile, which is measured in Purchasing Power Standard (PPS) per adult equivalent. This dataset reflects the relative magnitude of consumption expenditure by each income quintile and serves to compute the value of expenditure by each quintile. For example, Table 3 shows the consumption expenditure in PPS for Austria in 2015.

GEO (Labels)	Austria
QUANTILE (Labels)	
First quintile	17,096
Second quintile	20,274
Third quintile	23,350
Fourth quintile	23,639
Fifth quintile	28,874

Table 3: Consumption Expenditure by Quintiles

The fourth dataset, VAT rate by spending categories, is also retrieved from Eurostat – VAT rate by COICOP consumption purpose - experimental statistics. It provides annual VAT rate data by types of goods and services, measured in percentage. This dataset is used in calculation of the VAT burden. An example of Austria's VAT rate data is displayed in Table 4, where the COICOP labels match with Table 1.

GEO (Labels)	Austria
COICOP (Labels)	
Food and non-alcoholic beverages	10
Alcoholic beverages, tobacco and narcotics	20
Clothing and footwear	20
Housing, water, electricity, gas and other fuels	5.5
Furnishings, household equipment and routine household maintenance	20
Health	15.3
Transport	19.4
Communications	20
Recreation and culture	18.3

Education	0
Restaurants and hotels	19.3
Miscellaneous goods and services	17.2

Table 4: VAT Rate by Spending Categories

The fifth dataset, disposable income per capita, is from OECD – Households and NPISHs net adjusted disposable income per capita, measured in USD per person, Purchasing Power Parity (PPP) converted, 2021. It provides the value of disposable income per capita in the chosen reference area in a given year. For example, Austria’s disposable income per capita during the period 2015 to 2022 is shown in Table 5. The data would be used for assessing VAT’s impact on income distribution.

Time Period	2015	2016	2017	2018	2019	2020	2021	2022
Austria	41,336	41,874	42,252	42,577	42,837	41,473	42,781	43,452

Table 5: Disposable Income per capita (USD, PPP converted, 2021)

The sixth dataset, income share across quintiles, is obtained from Eurostat – Distribution of income by quantiles, measured as the share of national equivalised income. It serves the purpose of finding the total disposable income of each quintile before consumption and paying VAT. Table 6 is an example of Austria’s income distribution in 2015.

GEO (Labels)	Austria
QUANTILE (Labels)	
First quintile	9.0
Second quintile	14.1
Third quintile	18.0
Fourth quintile	22.5
Fifth quintile	36.3

Table 6: Income Share Across Quintiles

Besides, there are two other pieces of information collected to convert the datasets into the same currency. The average exchange rate between USD and EUR each year is from the European Central Bank, which would be used to adapt the currency to Euros for all data. In addition, the total population each year is obtained from the World Bank, altering all per capita data into aggregate data.

To assess the redistributive effects of VAT, additional processing is required for the consumption data. Using the first PPS consumption dataset, the proportion of each type of good in the total expenditure was calculated for each quintile. The second dataset provided the value of national total consumption in 2015 USD, which could be converted to euros using the exchange rate information. Combining with the data in the third dataset covering consumption expenditure by income quintile, the value of consumption expenditure by each quintile could be computed. Incorporating with the first dataset, the specific value of expenditure on each type of good by each quintile was calculated.

By matching the spending with VAT rates in the fourth dataset, the aggregate amount of VAT paid by each quintile could be calculated:

$$VAT\ Paid_q = \sum_i (Expenditure_{q,i} \times VAT\ Rate_i)$$

where q is the income quintile and i represents different types of goods and services.

To reach the conclusion on income distribution, the income data in the fifth dataset was adjusted using information on exchange rate and population. This resulted in the total disposable income in the country in a given year, which was integrated with income share across quintiles in the sixth dataset to find the total disposable income of each quintile before consumption and VAT.

By deducting the VAT paid from the disposable income, the disposable income after consumption and VAT (income after) can be calculated as:

$$Income\ After_q = Income\ Before_q - Total\ Expenditure_q - VAT\ Paid_q$$

As a result, the post-tax income shares of each quintile were then computed and compared to the pre-tax income shares. As shown in Table 7: The income share of the lowest income quintile (Q1) saw a drop from 9.00% to 7.45%, and the second and third quintiles also experienced a slight reduction in their income share, indicating that VAT places a higher burden on lower-income households. The fourth quintile experienced a minor rise in its income share, from 22.50% to 22.95%, while the highest income quintile (Q5) saw its income share increase from 36.30% to 39.13%, suggesting that VAT may further worsen inequality in an economy. The simulation with data from Austria in 2015 reveals a regressive VAT, thus proving that Hypothesis 1 is correct.

QUANTILE	INCOME SHARE BEFORE VAT	INCOME SHARE AFTER VAT
1	9.00%	7.45%
2	14.10%	13.14%
3	18.00%	17.33%
4	22.50%	22.95%

5	36.30%	39.13%

Table 7: Income share before and after VAT, Austria 2015

To further ensure the reliability of this conclusion, the study followed a similar process to calculate the income share before and after VAT and consumption for Belgium, also in the year 2015.

QUANTILE	INCOME SHARE BEFORE VAT	INCOME SHARE AFTER VAT
1	9.20%	0.13%
2	14.00%	5.73%
3	18.40%	14.74%
4	23.20%	24.64%
5	35.20%	54.75%

Table 8: Income share before and after VAT, Belgium 2015

The study obtained a more regressive result, shown in Table 8. The lowest income quintile (Q1) experienced a drastic drop in income share from 9.20 % to 0.13%, while the income share of the highest income quintile (Q5) rose by nearly 20% from 35.20% to 54.75%.

The data simulation with these two countries provided a very regressive view of VAT, aligning with the study's Hypothesis 1.

Regression

Through Hypothesis 2, the study aims to reassess the redistributive effects of VAT, covering 24 European countries over the time period 2015 to 2023. Europe was chosen for its wide adoption of VAT and the distinct rates in different countries, which could provide different perspectives on the topic. The study utilizes six datasets that include key variables relevant to VAT and income inequality, including the GINI coefficient, the dependent variable. The independent variables in this analysis are inflation rate, annual GDP, median and mean equivalised net income, and electoral democracy index.

The first dataset, inflation rate, is from the World Bank, under World Development Indicators. It includes the annual inflation rate of 266 countries across the globe, the earliest starting from 1960. Six countries are shown in Table 9, where the inflation rate is displayed in decimals.

Country Name	2015	2016	2017	2018	2019	2020	2021	2022	2023
Iraq	1.39	0.557	0.184	0.367	-0.199	0.574	6.04	4.99	4.36
Iceland	1.633	1.70	1.76	2.68	3.01	2.85	4.44	8.31	8.74

Israel	-0.601	-0.552	0.259	0.794	0.850	-0.614	1.51	4.39	4.23
Italy	0.039	-0.094	1.23	1.14	0.611	-0.138	1.87	8.20	5.62
Jamaica	3.69	2.35	4.38	3.74	3.91	5.23	5.86	10.3	6.47
Japan	0.795	-0.127	0.484	0.989	0.469	-0.0250	-0.233	2.50	3.27

Table 9: Inflation Rate

The second dataset, the electoral democracy index, is from Our World in Data. It indicates the extent to which political leaders are elected in free and fair elections and reflects freedom of association and expression. For example, the electoral democracy indices for Sweden in the time period 2015 to 2023 are laid out in Graph 8, where the indices range from 0 to 1.

Year	Electoral Democracy Index
2015	0.915
2016	0.914
2017	0.914
2018	0.909
2019	0.901
2020	0.901
2021	0.901
2022	0.896
2023	0.884

Table 10: Electoral Democracy Index for Sweden

The third dataset, GINI coefficient, is retrieved from Eurostat – Gini coefficient of equivalised disposable income by age. It measures annual Gini coefficients on a scale from 0 to 100. A part of the data set is displayed in Table 10, where each country has a corresponding Gini value in the years from 2015 to 2023.

TIME	2015	2016	2017	2018	2019	2020	2021	2022	2023
GEO (Labels)									
Belgium	26.2	26.3	26.1	25.7	25.1	25.4	24.1	24.9	24.2
Bulgaria	37	37.7	40.2	39.6	40.8	40	39.7	38.4	37.2

Czechia	25	25.1	24.5	24	24	24.2	24.8	24.8	24.4
Denmark	27.4	27.7	27.6	27.8	27.5	27.3	27	27.7	28.2
Germany	30.1	29.5	29.1	31.1	29.7	30.5	31.2	29	29.4
Estonia	34.8	32.7	31.6	30.6	30.5	30.5	30.6	31.9	31.8
Ireland	29.7	29.6	30.6	28.9	28	27.6	26.6	26.9	27.4
Greece	34.2	34.3	33.4	32.3	31	31.4	32.4	31.4	31.8
Spain	34.6	34.5	34.1	33.2	33	32.1	33	32	31.5
France	29.2	29.3	28.8	28.5	29.2	29.2	29.3	29.8	29.7

Table 11: Gini Coefficient 2015-2023

The three other datasets – annual real GDP per capita, median equivalised net income, and mean equivalised income – are all retrieved from Eurostat.

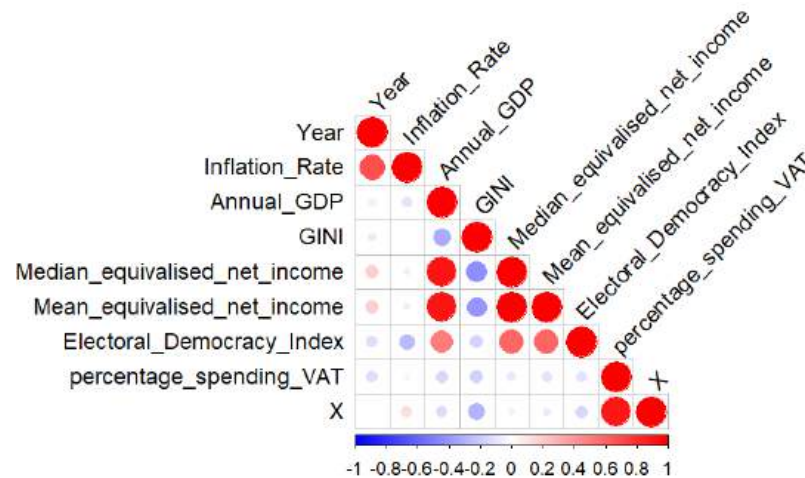
To join these datasets into a single table for analysis, further steps were needed. Due to the different sources and scopes of the datasets, the countries in each dataset were matched to ensure data consistency. In this case, only European countries were the subjects of the study, so the irrelevant information was cleared. An example of the cleaned dataset is laid out in Table 12, where the table for Belgium was displayed.

Year	Inflation Rate	Annual GDP	GINI	Median Equivalised Net Income	Mean Equivalised Net Income	Electoral Democracy Index	Percentage Spending VAT
2015	0.56%	36860	26.2	21654	23673	0.902	0.12591
2016	1.97%	37110	26.3	22295	24264	0.902	0.12591
2017	2.13%	37510	26.1	22733	24400	0.901	0.12591
2018	2.05%	38040	25.7	23686	25196	0.901	0.12591
2019	1.44%	38760	25.1	24608	26275	0.9	0.12591
2020	0.74%	36740	25.4	25672	27641	0.894	0.10556
2021	2.44%	38860	24.1	25855	27718	0.895	0.10556
2022	9.60%	40180	24.9	27314	29471	0.895	0.10556
2023	4.05%	40340	24.2	28997	31150	0.895	0.10556

Table 12: Integrated data for Belgium

The study runs a regression model to explore the relationship between the GINI coefficient and the independent variables. Before running the regression model, the correlation

matrix was reviewed to assess the relationship between the independent variables and eliminate potential multicollinearity.



Graph 1: Correlation between variables

Among all the independent variables, the correlation indices between annual GDP, median equivalised net income, and mean equivalised net income exceed 0.9, suggesting that only one of them could be in the regression model at a time. The correlation between any other pair of variables is below 0.6, which indicates low multicollinearity among the variables.

The Model 1 incorporates the inflation rate, percentage of spending subject to VAT, and annual GDP.

$$GINI = \beta_0 + \beta_1(\text{percentage spending VAT}) + \beta_2(\text{Inflation Rate}) + \beta_3(\text{Annual GDP}) + \varepsilon$$

VARIABLES	GINI	Standard Errors
Percentage_Spending_VAT	-26.36***	7.492
Inflation_Rate	1.635e-02	6.177e-02
Annual_GDP	-6.673e-05***	1.183e-05

Table 13: Regression Model 1; *** p<0.001, ** p<0.01, * p<0.05

The percentage of spending subject to VAT has a -26.36 coefficient, indicating that an increase in spending subject to VAT is linked to lower inequality. Annual real GDP is significantly associated with lower GINI, meaning that rises in GDP are linked to a fall in inequality. However, the association between the inflation rate and the GINI coefficient is not significant, showing that it has a negligible impact on inequality in this model. Besides, the model only explains 15.61% of the variation in the GINI coefficient, a slightly low number, suggesting that there are more relevant factors not included in this model.

In the Model 2, annual GDP is replaced with mean equivalised net income.

$$GINI = \beta_0 + \beta_1(\text{percentage spending VAT}) + \beta_2(\text{Inflation Rate}) + \beta_3(\text{Mean equivalised net income}) + \varepsilon$$

VARIABLES	GINI	Standard Errors
percentage_spending_VAT	-25.22***	7.216
Inflation_Rate	-5.312e-03	5.963e-02
Mean_equivalised_net_income	-1.549e-04***	2.238e-05

Table 14: Regression Model 2

The model explains 20.82% of the variation in the GINI coefficient, an improvement compared to Model 1. The percentage of spending subject to VAT retains its significant and negative effect, and the inflation rate remains insignificant. On the other hand, mean equivalised net income has a highly significant negative effect, suggesting that higher average income is associated with lower inequality.

This model provides a slightly better explanation of the variation in GINI. The new variable mean income is confirmed to be highly significant. At the same time, more approaches could be taken to perfect the model, such as replacing inflation rate, which shows little explanatory power in this model.

The final model, Model 3, replaces the inflation rate with the democracy index and applies a logarithm on mean income to better fit the model.

$$GINI = \beta_0 + \beta_1(\text{percentage spending VAT}) + \beta_2(\log(\text{Mean equivalised net income})) + \beta_3(\text{Electoral+Democracy Index}) + \varepsilon$$

VARIABLES	GINI	Standard Errors
percentage_spending_VAT	-23.1602***	6.8136
log(Mean_equivalised_net_income)	-3.8161***	0.4608
Electoral_Democracy_Index	9.0400**	2.9440

Table 15: Regression Model 3

As shown in Table 15, the percentage of spending subject to VAT remains negative and significant, indicating that a higher share of spending subject to VAT is associated with lower inequality. The log-transformed mean income has a higher absolute value of the coefficient, suggesting that income level has a stronger association with inequality when viewed proportionally. The new variable, the electoral democracy index, has a positive and significant coefficient, showing that democracy is positively linked with inequality.

The model explains 29.27% of the variation in GINI, which suggests a substantial explanation for inequality. Hence, based on Model 3 and the results from the two previous

models, it could be concluded that across all models, a higher percentage of spending subject to VAT is associated with lower inequality. In other words, VAT is a progressive tax, proving Hypothesis 2.

Conclusion

The simulation and regression analyses indicate that VAT is generally associated with inequality. The simulations for Belgium and Austria suggest regressive short-term effects, whereas regression results show a negative relationship between VAT coverage and inequality. Model 3, explaining 29.27% of the variation in GINI, estimates that a 1% increase in the share of total spending subject to VAT is correlated with a 23.16% decrease in the GINI coefficient. These findings suggest that VAT, contrary to conventional belief, has a progressive effect by promoting a more equal income distribution. The study reviewed prior studies from differing methodological frameworks to provide a comprehensive theoretical basis. Data were drawn primarily from Eurostat, ensuring consistency and reliability for cross-country analysis, and missing data were supplemented with verified national sources. Country-specific VAT policies, including reductions and exemptions, were incorporated, enhancing real-world relevance. The analysis also employed multiple approaches: an initial simulation model confirmed the traditional regressive view, while a subsequent regression model which accounted for multiple economic variables indicated that VAT is associated with reduced inequality.

Future studies could extend the temporal and geographical scope of the data to add more generalizability of the results. We also suggest that future research could employ more advanced models to offer deeper insights. Overall, the study contributes meaningful evidence that challenges the conventional assumption of VAT's regressivity. Its findings hold potential implications for policymakers considering VAT reforms or shifts toward indirect taxation. Further research should extend the dataset across more countries and longer time periods and integrate advanced econometric techniques to better assess the long-term and structural effects of VAT on inequality.

Works Cited

- Tait, Alan A. Value Added Tax International Practice and Problems. Washington, D.C. International Monetary Fund, 1988.
- Pechman, Joseph A, and Benjamin A Okner. Who Bears the Tax Burden? Washington : Brookings Institution, 1974.
- Cathal O'Donoghue, et al. "Modelling the Redistributive Impact of Indirect Taxes in Europe: An Application of EUROMOD." RePEc: Research Papers in Economics, 1 June 2004. Accessed 7 Feb. 2025.
- CASPERSEN, ERIK, and GILBERT METCALF. "IS A VALUE ADDED TAX REGRESSIVE? ANNUAL versus LIFETIME INCIDENCE MEASURES." National Tax Journal, vol. 47, no. 4, 1 Dec. 1994, pp. 731–746, <https://doi.org/10.1086/ntj41789105>. Accessed 19 Mar. 2021.
- Arsić, Miloško, and Nikola Altiparmakov. "Equity Aspects of VAT in Emerging European Countries: A Case Study of Serbia." Economic Systems, vol. 37, no. 2, June 2013, pp. 171–186, <https://doi.org/10.1016/j.ecosys.2012.12.003>. Accessed 22 Mar. 2022.
- Decoster, André, et al. "How Regressive Are Indirect Taxes? A Microsimulation Analysis for Five European Countries." Journal of Policy Analysis and Management, vol. 29, no. 2, Mar. 2010, pp. 326–350, <https://doi.org/10.1002/pam.20494>. Accessed 12 Dec. 2019.
- Allison, Paul D. "Measures of inequality." American sociological review (1978): 865-880.
- Dalton, Hugh. "The Measurement of the Inequality of Incomes." The Economic Journal, vol. 30, no. 119, Sept. 1920, p. 348, www.ophi.org.uk/wp-content/uploads/Dalton-1920.pdf, <https://doi.org/10.2307/2223525>.
- Cowell, Frank. "Measuring Inequality." Journal of Economics, vol. 104, no. 3, 29 Apr. 2011, pp. 281–285, <https://doi.org/10.1007/s00712-011-0211-2>.
- Gatzia, Dimitria E, and Douglas Woods. "Progressive Taxation as a Means to Equality of Condition and Poverty Alleviation." Economics Management and Financial Markets, vol. 9, no. 4, 1 Dec. 2014, p. 29. Accessed 8 Feb. 2025.
- "Rethinking Tax: The Shift to Indirect Tax." Grant Thornton International Ltd. Home, 7 Mar. 2016, www.grantthornton.global/en/insights/articles/rethinking-tax-the-shift-to-indirect-tax/. Accessed 20 Mar. 2025.

Implications of Genetic, Environmental, and Immunological Factors in Oral Lichens Planus and Their Insight on Therapeutic Interventions By Anshika Alamuthu

Abstract

Oral Lichen Planus (OLP) is a chronic inflammatory disorder of the oral mucosa with a multifactorial etiology involving genetic predisposition, environmental triggers, and dysregulated immune responses. Affecting approximately 1–2% of the global population, OLP can significantly impair quality of life due to chronic pain, mucosal damage, and potential malignant transformation. Clinically, OLP manifests in diverse forms, ranging from asymptomatic lesions to painful erosive presentations, posing diagnostic and therapeutic challenges. While significant research has elucidated key aspects of its pathogenesis, particularly the role of T cell-mediated cytotoxicity and inflammatory cytokines, the precise mechanisms driving disease progression remain incompletely understood. This review critically examines current evidence on the genetic, environmental, and immunological factors contributing to OLP, with particular focus on biomarkers of disease susceptibility and severity, and how these insights inform the development of targeted therapies such as immune-modulating agents and precision medicine strategies. By integrating knowledge across these domains, it aims to improve diagnostic accuracy, clinical management, and long-term treatment outcomes.

Introduction

Chronic inflammatory disorders of the oral cavity represent a significant burden on global health, impairing essential functions such as speaking, eating, and overall quality of life. Among these conditions, Oral Lichen Planus (OLP) has gained particular attention because of its multifactorial origins and its potential for malignant transformation. The condition illustrates how complex interactions between host genetics, environmental exposures, and immune dysregulation converge to produce persistent mucosal disease (Gonzalez-Moles et al., 2008; Sugerman et al., 2002).

OLP is now recognized as an immune-mediated disorder in which T-cell-driven inflammatory responses target the oral epithelium, producing chronic lesions. Clinically, OLP presents in several forms. The reticular type is marked by white, lace-like striations known as Wickham's striae. The erosive type involves painful ulcerations and erythema, while the atrophic type is characterized by mucosal thinning and increased sensitivity (Lodi et al 2005)

(Figure 1)

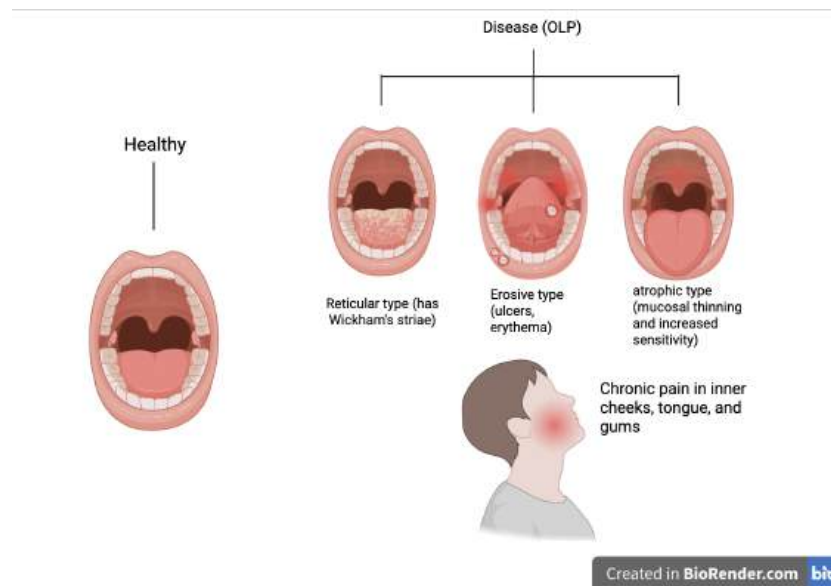


Figure 1: Symptoms in OLP Comparison of healthy oral mucosa and tissue affected by Oral Lichen Planus (OLP). OLP is characterized by chronic, immune-mediated inflammation resulting in oral lesions of varying clinical types—reticular (white striations), erosive (painful ulcerations), and atrophic (mucosal thinning). These manifestations contribute to hallmark symptoms such as chronic pain and oral sensitivity, as described in the clinical overview of OLP.

Diagnosis of Oral Lichen Planus (OLP) requires a comprehensive approach combining clinical evaluation with histopathological confirmation by dental, dermatological, or oral pathology specialists. Key histological features include hyperkeratosis, saw-toothed rete ridges, degeneration of the basal cell layer, and a dense, band-like subepithelial lymphocytic infiltrate (Van der Meij & van der Waal, 2003). Because of the potential for malignant transformation, with reported rates between 0.4 and 1.74 percent, biopsy is essential to exclude premalignant or malignant disease (Warnakulasuriya et al., 2010). In cases of diagnostic uncertainty, direct immunofluorescence can be used to confirm OLP and distinguish it from other lichenoid lesions such as lichenoid drug reactions, lupus erythematosus, and graft-versus-host disease, which may present overlapping histological features but require different management strategies (van der Meij & van der Waal, 2003; Al-Hashimi et al., 2007).

Management strategies generally focus on reducing inflammation, alleviating pain, and monitoring for malignant transformation. While corticosteroids and other immunomodulating agents remain the most common therapeutic approaches, their effectiveness can vary, and adverse effects or treatment resistance are not uncommon (Thongprasom et al., 2007; Gonzalez-Moles et al., 2013; Akram et al., 2018; Mico-Llorens et al., 2020). These challenges underscore the importance of advancing knowledge about the genetic, environmental, and immunological factors that drive OLP, since a deeper understanding of its pathogenesis is essential for improving both diagnosis and long-term management.

To build this understanding, it is essential to first examine the immunological mechanisms that underpin OLP, since immune dysfunction forms the biological foundation upon which genetic and environmental influences act.

Immunological Factors and Response

Immunologic dysregulation lies at the center of OLP pathogenesis. The disorder is classified as a T-cell-mediated autoimmune disease marked by chronic mucosal inflammation. In healthy tissue, inflammation is a protective process that eliminates pathogens and promotes repair. In OLP, however, this response becomes pathological as immune cells attack the body's own oral epithelium.

The immune cascade begins when antigen-presenting cells (APCs), including dendritic and Langerhans cells, detect abnormal antigens in the oral mucosa such as altered self-proteins, viral components, or drug metabolites. These antigens are processed and displayed on major histocompatibility-complex (MHC) class I and II molecules, which activate both CD8⁺ cytotoxic T-cells and CD4⁺ helper T-cells. Activated CD8⁺ cells trigger apoptosis of basal keratinocytes through granzyme B release and engagement of the Fas/FasL pathway (Sugerman et al., 2002; Zhou et al., 2002). Simultaneously, CD4⁺ helper T-cells secrete cytokines such as interleukin-2 (IL-2), interferon-gamma (IFN- γ), and tumor-necrosis-factor-alpha (TNF- α), which recruit additional inflammatory cells and amplify local damage (Rhodus et al., 2005; Pannunzio et al., 2010) (**Figure 2**).

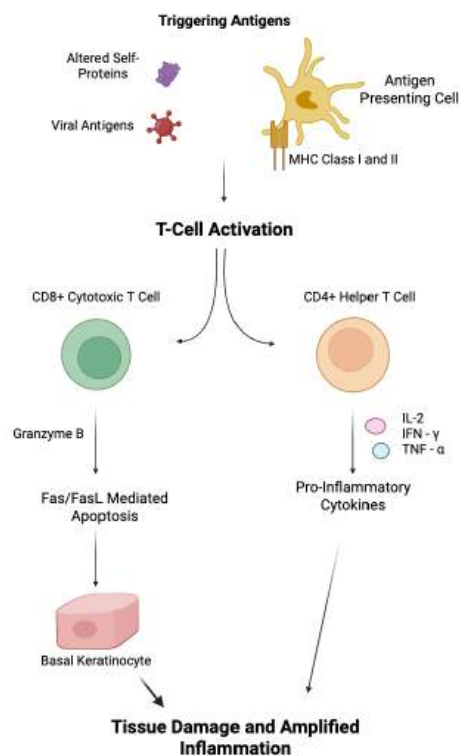


Figure 2: Immune Response in OLP Antigen-presenting cells (APCs), such as dendritic and Langerhans cells, initiate the immune response by detecting and processing abnormal antigens within the oral mucosa. These are

presented via MHC molecules to CD8+ and CD4+ T cells. CD8+ cytotoxic T cells induce apoptosis in basal keratinocytes through granzyme B and Fas/FasL pathways, while CD4+ helper T cells amplify local inflammation through cytokines like IL-2, IFN- γ , and TNF- α . This dual activation contributes to chronic mucosal inflammation and epithelial damage characteristic of OLP.

This immune process creates a self-perpetuating inflammatory loop in which continuous antigen presentation maintains T-cell activation and epithelial destruction. Each cycle of keratinocyte apoptosis releases new self-antigens that are reprocessed and presented again, sustaining chronicity. Among T-cell subsets, Th1 and Th17 cells are predominant. Th1 cells secrete IFN γ , enhancing macrophage activation and MHC expression, whereas Th17 cells release IL17, recruiting neutrophils that exacerbate tissue damage (Carrozzo & Scally, 2014; Zhang et al., 2014).

A reduction or functional impairment of regulatory T-cells (Tregs) worsens this imbalance. Tregs normally suppress autoreactive T-cells and maintain immune tolerance, but in OLP their dysfunction allows persistent effector-cell activation and chronic inflammation (Li et al., 2019; Yang et al., 2020). This immunologic pattern parallels other autoimmune diseases such as systemic lupus erythematosus and rheumatoid arthritis, indicating shared pathogenic pathways. Elevated levels of IL6, IL17, IFN γ , and TNF α have been identified in OLP lesions, contributing to oxidative stress and potentially increasing the risk of malignant transformation (Wang et al., 2021; Wei et al., 2023) (**Figure 3**).

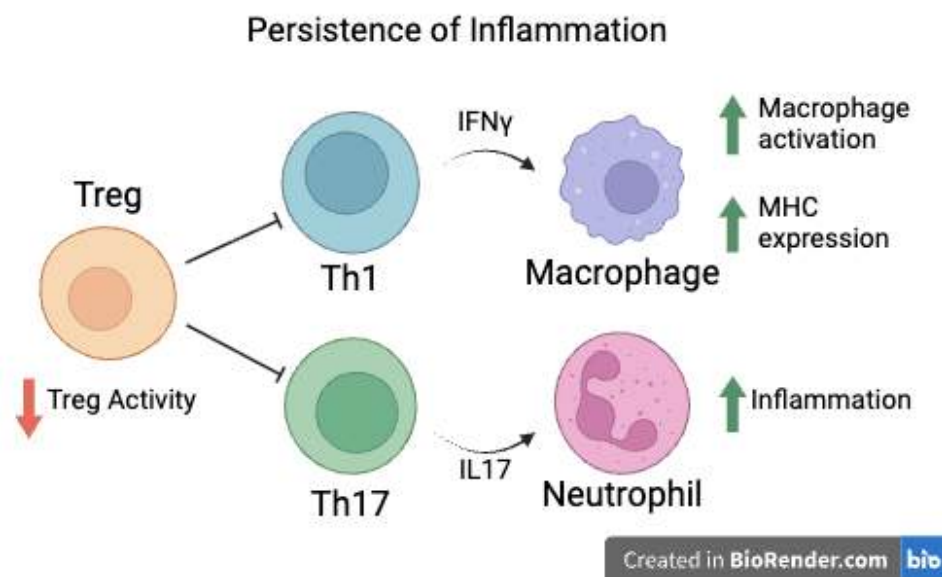


Figure 3: Persistence of Inflammation Decreased activity of Treg cells in OLP releases inhibition of Th1 and Th17 subsets of CD4+ T-cells. These activated Th1 and Th17 cells secrete IFN γ and IL17, respectively. IFN γ leads to increased macrophage activity and MHC expression. IL17 leads to increased inflammation through recruitment of neutrophils. Overall, this leads to tissue damage.

Histopathologic findings reflect these immune disturbances. A dense, band-like lymphocytic infiltrate forms along the epithelial–connective-tissue interface, accompanied by

basal-cell degeneration and the presence of Civatte bodies, which are apoptotic keratinocytes considered diagnostic of OLP (Lodi et al., 2005). Collectively, these observations show that OLP is sustained by an imbalance between hyperactive cytotoxic and helper T-cells and insufficient regulatory control. Recognizing this immune foundation clarifies how genetic and environmental factors further modify susceptibility and disease expression.

Environmental Triggers

Environmental influences can initiate or worsen Oral Lichen Planus (OLP) in individuals who already have genetic or immunologic susceptibility. These triggers include pharmacological, chemical, psychological, and infectious factors, each capable of disturbing immune balance and promoting mucosal inflammation. Understanding these categories helps explain how external exposures interact with intrinsic genetic and immunologic mechanisms to produce disease.

Pharmacological and Chemical Exposures

Medications are among the most frequently reported environmental contributors to OLP. Drugs such as angiotensin-converting-enzyme (ACE) inhibitors, thiazide diuretics, non-steroidal anti-inflammatory drugs (NSAIDs), antimalarials, antimicrobials, antihypertensives, psychiatric agents, antidiabetics, and immune-checkpoint inhibitors (for example PD-1 blockers used in oncology) have all been associated with lichenoid reactions (Ismail et al., 2007; Scully & Carrozzo, 2008). These compounds can modify keratinocyte antigens or generate reactive metabolites that act as haptens, which then stimulate a T-cell-mediated immune response. In addition to medications, dental materials are recognized chemical triggers. Alloys that contain mercury, nickel, or chromium can release ions that bind to host proteins and form hapten-protein complexes, activating local T-cells. The resulting localized inflammation around dental restorations may lead to clinically visible OLP lesions (Seymour et al., 1993; Garcia-Pola et al., 2002). Patch testing and replacement of amalgam restorations have been associated with symptom improvement in sensitized patients.

Ultraviolet (UV) filters and cosmetic additives, such as benzophenones and cinnamates found in sunscreens, lip balms, and hair products, can also provoke lichenoid eruptions. These chemicals can cause photo-allergic reactions, particularly in individuals with heightened cutaneous sensitivity. Discontinuation of such products has led to clinical improvement in up to 30 percent of reported cases (Pereira et al., 2012; García-Pola et al., 2010). Collectively, these observations show that chemical and pharmacologic exposures can both trigger and perpetuate OLP through immune-sensitization mechanisms.

Psychological and Physiological Stressors

Psychological stress is another significant environmental factor. Activation of the hypothalamic-pituitary-adrenal (HPA) axis during stress increases cortisol and catecholamine levels, which in turn alter cytokine production and weaken mucosal immune regulation. Chronic

stressors such as grief, occupational strain, or trauma are associated with flare-ups or new lesion development in OLP patients (Chaudhary, 2004).

Quantitative psychometric studies strengthen this connection. Rojo-Moreno et al. (2010) found that individuals with high anxiety or depression scores were two to three times more likely to experience disease exacerbations than controls. These findings suggest that psychosocial interventions, including stress-management programs and cognitive-behavioral therapy, could play a supportive role in OLP management by reducing immune dysregulation.

Infectious Agents

Chronic viral infections, particularly hepatitis C virus (HCV), show a strong association with OLP. Serologic studies reveal that anti-HCV antibodies are significantly more common in OLP patients than in the general population (Nagao et al., 2000). The leading explanation involves molecular mimicry, where viral antigens resemble host proteins and elicit cross-reactive T-cell responses that attack oral epithelial cells (Carrozzo & Scally, 2014).

In regions where HCV infection is endemic, the coexistence of OLP and liver disease has diagnostic value, and clinicians are encouraged to screen for viral markers in unexplained cases. Other infectious agents such as human papillomavirus (HPV), Epstein–Barr virus (EBV), and *Helicobacter pylori* have been suggested as possible cofactors, although definitive evidence remains limited. Overall, infection-driven immune activation can cooperate with genetic predisposition to sustain chronic mucosal inflammation.

Environmental triggers therefore act through overlapping pathways, including chemical sensitization, neuro-immune modulation, and pathogen-induced autoimmunity. Recognizing and minimizing these exposures remains a critical step in managing OLP effectively.

Genetic Factors

Genetic studies provide important insight into why certain individuals are predisposed to OLP. Susceptibility arises from a combination of inherited and epigenetic influences. One of the most consistent genetic associations involves variations in Human Leukocyte Antigen (HLA) alleles. HLA genes are part of the major histocompatibility complex (MHC) located on chromosome 6, which coordinates immune recognition by encoding cell-surface proteins that present antigenic peptides to T-cells (Shiina et al., 2009). When functioning normally, this process distinguishes self from non-self antigens, but specific HLA variants may predispose the immune system to misrecognize self-proteins and trigger autoimmunity.

Both class I (HLA-A, HLA-B, HLA-C) and class II (HLA-DR, HLA-DQ, HLA-DP) molecules are implicated. Class I molecules present endogenous antigens to CD8⁺ cytotoxic T-cells, while class II molecules present exogenous peptides to CD4⁺ helper T-cells (Neefjes et al., 2011). Studies reveal that HLA-DR and HLA-DQ alleles occur more frequently among OLP patients, especially in certain ethnic populations, implying a genetic predisposition to dysregulated T-cell activation (Dennis, 2012; López-Jornet et al., 2012).

Beyond HLA variation, other genetic factors contribute. Deficiencies in the vitamin D pathway, which normally promotes immune tolerance and inhibits pro-inflammatory cytokines, have been associated with enhanced OLP severity. Polymorphisms in cytokine genes, including IL-6 and IL-10, are also linked to altered immune responses. These variations can increase pro-inflammatory signaling or reduce anti-inflammatory feedback, sustaining the chronic inflammation that defines OLP (Bazrafshani et al., 2010).

Epigenetic modifications add another regulatory layer. They involve heritable but reversible changes in gene expression without DNA-sequence alteration. DNA methylation and histone modification are the principal mechanisms. DNA methylation usually silences genes by adding methyl groups to cytosine residues within CpG islands, a process mediated by DNA methyltransferases (DNMTs) (Shavit et al., 2020). In OLP, hypermethylation of E-cadherin and p16^{INK4a} has been reported. Reduced E-cadherin expression may weaken epithelial adhesion and promote inflammation, whereas suppression of p16^{INK4a}, a tumor-suppressor gene, may allow uncontrolled cell proliferation and increase malignant potential.

Histone modifications, including acetylation, methylation, and phosphorylation, also influence gene accessibility. Global hyperacetylation of histone H3 has been documented in T-cells from OLP patients, which may decrease expression of anti-inflammatory genes and maintain immune activation. Together, these genetic and epigenetic findings explain how immune tolerance can be lost and chronic inflammation maintained in susceptible individuals.

Therapies

Management of OLP primarily aims to relieve symptoms, control inflammation, and prevent malignant transformation rather than achieve a complete cure. Current therapeutic options include topical and systemic medications, biologic agents, photodynamic therapy, and complementary approaches.

Conventional Anti-inflammatory Treatments

Topical corticosteroids remain the first-line therapy because of their strong anti-inflammatory effects and localized action. Agents such as clobetasol propionate and fluocinonide inhibit T-cell activation and cytokine production, producing symptomatic relief in about 70 to 80 percent of patients (Edwards & Kelsch, 2002; Thongprasom et al., 2007). Long-term use can cause mucosal thinning, oral candidiasis, and decreased drug response. Systemic corticosteroids are reserved for severe or widespread cases but must be used cautiously due to potential systemic side effects including immunosuppression, hypertension, and weight gain.

Calcineurin Inhibitors and Immunomodulators

For steroid-resistant patients, calcineurin inhibitors such as tacrolimus and pimecrolimus offer an alternative. These drugs block IL-2 transcription and suppress T-cell activation. Clinical studies have shown lesion regression in 70 to 82 percent of patients treated with topical

tacrolimus, though some experience burning sensations or require close monitoring for potential carcinogenic risk (Lodi et al., 2005; Thongprasom et al., 2007).

Biologic and Targeted Therapies

Recent advances in immunology have led to the exploration of biologic agents that specifically target cytokines involved in OLP. Tumor-necrosis-factor-alpha (TNF- α) inhibitors such as etanercept and infliximab, and interleukin-17 (IL-17) blockers such as secukinumab, have shown encouraging preliminary results. Case reports describe symptom reduction and lesion healing (Chaudhary et al., 2013; Zhang et al., 2020). However, cost, limited sample sizes, and safety concerns prevent routine clinical use. Large randomized studies are required to establish efficacy.

Photodynamic Therapy (PDT)

Photodynamic therapy is a promising non-invasive method for refractory OLP. It uses a photosensitizing agent, commonly methylene blue or 5-aminolevulinic acid, followed by activation with light of a specific wavelength to produce reactive oxygen species that destroy inflammatory and dysplastic cells. Meta-analyses indicate that PDT achieves 60 to 70 percent lesion reduction and significant pain relief with minimal adverse effects (Akram et al., 2018; Ghahremanloo et al., 2022). Because of its localized action and low toxicity, PDT represents an appealing long-term option for patients with chronic disease.

Natural and Adjunctive Therapies

Complementary treatments such as curcumin, aloe vera, and purslane have also been studied for their antioxidant and anti-inflammatory effects. Curcumin has produced 50 to 60 percent reduction in lesion severity in small clinical trials (Chainani-Wu et al., 2007; Mico-Llorens et al., 2020). Aloe vera gels and purslane extracts have shown moderate improvements in pain and healing (Gottumukkala et al., 2013; Nafiye et al., 2019). However, differences in formulation and dosage among studies make results difficult to generalize. Emerging research on immunogenetic and epigenetic alterations in OLP, such as FOXP3+ regulatory T-cell dysfunction and abnormal DNA methylation, suggests potential for precision-medicine approaches. Future treatments might include targeted biologics or small molecules tailored to individual immune profiles, although these remain conceptual rather than clinical (Li et al., 2019; Yang et al., 2020).

Discussion

The current understanding of Oral Lichen Planus (OLP) reveals significant strides in elucidating its immunopathological mechanisms, yet many questions remain unresolved. As discussed earlier, the persistence of OLP is largely due to sustained T-cell-mediated inflammation and inadequate regulatory immune responses. While these insights have supported the continued use of immunomodulatory therapies such as corticosteroids and calcineurin

inhibitors, the field lacks consensus on long-term management, particularly for refractory or relapsing cases.

One of the most promising areas of recent exploration is the introduction of biologic therapies targeting cytokines such as TNF- α and IL-17. Although these agents—originally developed for other autoimmune diseases like psoriasis and rheumatoid arthritis—have shown early potential in OLP, their use remains limited to preclinical studies and small Phase I/II clinical trials (González-Moles et al., 2021). Concerns over cost, immune suppression, and long-term safety have slowed their integration into standard treatment regimens.

Photodynamic therapy (PDT) has also gained attention as a non-invasive alternative with minimal systemic effects. Early studies report a 60–70% reduction in lesion size and pain (Akram et al., 2018), yet larger, multicenter trials are needed to validate these findings and establish standardized protocols. Similarly, natural agents such as curcumin and aloe vera have demonstrated modest anti-inflammatory benefits in small trials (Mico-Llorens et al., 2020), but variability in study design and dosage weakens the strength of current evidence.

In recent years, the identification of biomarkers has emerged as a critical focus in understanding disease progression, diagnosis, and prognosis. Biomarkers such as inflammatory cytokines, salivary metabolites, proteomic signatures, and noncoding RNAs have shown potential to bridge the gap between molecular mechanisms and clinical application. Studies have demonstrated that elevated levels of IFN- γ , IL-17, and TNF- α in tissue and saliva correlate with disease severity and activity, while metabolomic profiling has identified altered salivary compounds like methylamine and tyrosine that differentiate OLP from healthy controls with high accuracy (Kashyap et al., 2024). Furthermore, a predictive diagnostic model based on immune-related genes such as BST2, RNASEL, PI3, DEFB4A, and CX3CL1 has shown strong diagnostic performance, underscoring the potential for molecular biomarkers to assist in early detection and personalized therapy (Bian et al., 2024). Collectively, these studies indicate that biomarkers could become valuable tools for stratifying patients by risk, monitoring treatment response, and identifying early malignant transformation.

However, despite encouraging results, several challenges remain before biomarker integration into clinical practice becomes feasible. Many studies are limited by small sample sizes, lack of standardized protocols, and insufficient longitudinal validation. Chen et al. (2024) emphasized that cytokine-related markers may not only reflect immune activation but also play a causal role in disease progression, suggesting that molecular pathways could eventually guide precision-targeted therapies. To advance this field, multicenter prospective trials and unified diagnostic criteria are needed to validate sensitivity, specificity, and reproducibility across populations.

Despite OLP's potential for malignant transformation and its chronic impact on quality of life, research into novel, targeted therapies remains sparse compared to other mucocutaneous disorders. This disparity reflects both the under-recognition of OLP and the broader challenges in studying rare or complex inflammatory diseases. Moving forward, future studies should

prioritize the development of robust clinical trials, validated diagnostic biomarkers, and more nuanced classifications of disease subtypes.

In conclusion, while current treatment strategies for OLP aim to manage symptoms and reduce inflammation, they do not address the underlying disease mechanisms or prevent progression. The expanding field of biomarker research offers a promising bridge between immunopathology and clinical translation, with the potential to improve diagnostic accuracy, monitor disease activity, and identify high-risk cases. Continued investigation into the molecular and immunological drivers of OLP is essential to shift from palliative care toward disease-modifying, precision-guided interventions.

Acknowledgements

The author thanks Miranda L. for her mentorship, guidance, and expertise in the creation of this research review, as well as critical reading of this manuscript.

Works Cited

- Abdeldayem, M., et al. (2022). Salivary expression of IL-17 and long noncoding RNA DQ786243 in oral lichen planus. *Journal of Oral Pathology & Medicine*, 51(3), 250–258. <https://doi.org/10.1111/jop.13287>
- Akram, Z., Javed, F., Hosein, M., & Almas, K. (2018). Photodynamic therapy as an anticancer approach for the treatment of oral lichen planus: A systematic review. *Photodiagnosis and Photodynamic Therapy*, 21, 147–153. <https://doi.org/10.1016/j.pdpdt.2017.11.018>
- Al-Hashimi, I., Schifter, M., Lockhart, P. B., Wray, D., Brennan, M., Migliorati, C. A., ... & Holmstrup, P. (2007). Oral lichen planus and oral lichenoid lesions: Diagnostic and therapeutic considerations. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology*, 103(S1), S25.e1–S25.e12. <https://doi.org/10.1016/j.tripleo.2006.11.034>
- Al-Jamaei, A., et al. (2022). Immunohistochemical markers in oral lichen planus: A systematic review. *Oral Diseases*, 28(2), 309–321. <https://doi.org/10.1111/odi.13982>
- Bazrafshani, M. R., et al. (2010). TGF- β 1, IL-10, IL-6 and TNF- α gene polymorphisms in Iranian patients with oral lichen planus. *Journal of Oral Pathology & Medicine*, 39(9), 730–734. <https://doi.org/10.1111/j.1600-0714.2010.00926.x>
- Bian, J., et al. (2024). Development of an immune-related diagnostic predictive model for oral lichen planus. *Medicine*, 103(11), e37469. <https://doi.org/10.1097/MD.00000000000037469>
- Carrozzo, M., & Scally, K. (2014). Oral manifestations of hepatitis C virus infection. *World Journal of Gastroenterology*, 20(24), 7534–7543. <https://doi.org/10.3748/wjg.v20.i24.7534>
- Chen, X., et al. (2024). Inflammatory cytokines and oral lichen planus: A Mendelian randomization study. *Frontiers in Immunology*. <https://doi.org/10.3389/fimmu.2024.1332317>
- Dennis, D. (2012). Genetic aspects of oral lichen planus: A review. *International Journal of Contemporary Dentistry*, 3(2), 1–5.
- Edwards, P. C., & Kelsch, R. (2002). Oral lichen planus: Clinical presentation and management. *Journal of the Canadian Dental Association*, 68(8), 494–499.
- Gonzalez-Moles, M. A., et al. (2021). Immune-mediated pathogenesis and emerging therapies in oral lichen planus. *Frontiers in Oral Health*, 2, 756. <https://doi.org/10.3389/froh.2021.756>
- Kashyap, B., et al. (2024). Salivary metabolomics in patients with oral lichen planus. *Clinical Oral Investigations*. <https://doi.org/10.1007/s00784-023-05389-1>
- Li, D., Liu, Y., Li, Y., & Wu, J. (2019). Imbalance between Th17 and regulatory T cells in patients with oral lichen planus. *Mediators of Inflammation*, 2019, 7154871. <https://doi.org/10.1155/2019/7154871>
- Lodi, G., Scully, C., Carrozzo, M., Griffiths, M., Sugerman, P. B., & Thongprasom, K. (2005). Current controversies in oral lichen planus: Report of an international consensus meeting.

Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology, 100(2), 164–178. <https://doi.org/10.1016/j.tripleo.2004.06.076>

Mico-Llorens, J. M., Chimenos-Küstner, E., Estrugo-Devesa, A., Jané-Salas, E., & López-López, J. (2020). Efficacy of topical treatments for the management of symptomatic oral lichen planus: A systematic review and network meta-analysis. *Journal of Clinical Medicine*, 9(5), 1343. <https://doi.org/10.3390/jcm9051343>

Wang, Y., et al. (2021). Up-regulated IL-17 and IL-23 in the immune pathogenesis of oral lichen planus. *Mediators of Inflammation*, 2021, 6636641. <https://doi.org/10.1155/2021/6636641>

Warnakulasuriya, S., Johnson, N. W., & van der Waal, I. (2010). Nomenclature and classification of potentially malignant disorders of the oral mucosa. *Journal of Oral Pathology & Medicine*, 36(10), 575–580. <https://doi.org/10.1111/j.1600-0714.2007.00582.x>

Biological and Clinical Effects of Genetic Mutations in Female Associated Cancers

By Mira Rakesh

ABSTRACT

This paper explores the relationship between genetic mutations and female cancer development, focusing on the BRCA1/2, PTEN, and ATM gene mutations. BRCA1/2, PTEN, and ATM are tumor suppressor genes, meaning they regulate the cell cycle and prevent uncontrolled cell growth. Cancer occurs when cell growth is unrestrained, forming a clump of cells called a tumor. When functioning properly, BRCA1/2, PTEN, and ATM genes are able to stop cancer cells from spreading and block tumor formation. When there is an inherited or somatic mutation in the gene's genetic material, they are not able to perform regular cell division and repair DNA damage, contributing to the spread of cancer. Being a holder of one of these gene mutations significantly increases the risk of developing female specific cancers like breast cancer, ovarian cancer, and endometrial cancer. By providing an overview of the function of these genes and the cellular effects of their mutations in relation to cancer development, this review aims to provide insight into the clinical implications, diagnosing, and treatment of cancer and cancer-causing gene mutations.

INTRODUCTION

Cancer is a semblance of disfunctioning cells that often appear as a tumour. Cancer occurs when regular cells turn malignant, a change caused by mutations. Mutations can be hereditary (inherited from our parents) or somatic (developed after fertilization). (Brown et al., 2023). Mutations occur when the DNA sequence of a gene is changed (Brown, 2002). Cancer causing mutations often occur in genes that are vital for controlling the cell cycle (cell proliferation and gamete production) (Collins et al., 1997). When these genes are mutated, they cannot function properly. In a normal cell, telomeres (ends of chromosomes) shorten after each cell division. When the telomere gets too short the cell goes to senescence (essentially cell death). A cancer cell fights against this function by activating telomerase to rebuild the cell's telomeres, allowing cancer cells to continue to proliferate forever (Deng et al., 2008). In addition to this, cancer cells require high levels of nutrients and oxygen to maintain the rate at which they proliferate (Liberti & Locasale, 2017). Cancer cells require an increased uptake of glucose to compensate for its high levels of cell division and growth. Because of this, the cells rely on aerobic glycolysis, the process of converting glucose into lactate known as the Warburg effect (Vander Heiden et al., 2009).

In positron emission tomography (PET imaging), patients are injected with glucose analogue FDG. FDG cannot be metabolized and gets trapped inside cells with high glucose uptake (cancer cells) (Gambhir, 2002). As FDG decays, it emits signals for the PET scanner to detect, showing bright spots where cancer cells are clustered (Bose et al., 2021). Magnetic resonance imaging (MRI) is a non-invasive method that does not use radioactive tracers, instead uses a strong magnetic field and radio waves to create a detailed image of body tissues (Philip,

2003). The MRI aligns hydrogen nuclei (protons) in the body with a powerful magnet. Then, the MRI scanner sends radio pulse waves to temporarily unalign the protons. When the radio wave stops, the protons move to their regular alignment and emit tiny radio signals (Lugauer & Wetzl, 2018). The MRI machine detects these signals and uses them to create pictures of the tissues, where tumors can be easily found because of their many different properties such as different cellular density and surface area, and abnormal blood flow (Lugauer & Wetzl, 2018)(Baba & Câtoi, 2007).

Proteins are created from a cell's genetic material (DNA) in protein synthesis. In protein synthesis, DNA is first converted to RNA through transcription in a process where DNA is separated into two single strands and then transcribed into RNA. The RNA strand is then further modified into mRNA (messenger RNA). Ribosomes then bind to mRNA and recruit tRNA (transfer RNA) molecules to bring amino acids corresponding to the mRNA codons (groups of 3 ribonucleic acid molecules). The ribosome binds these amino acids together as it reads further down the mRNA strand, creating a polypeptide chain and eventually a full protein (Alberts et al., 2002).

There are different types of mutations, all stemming from a change in the DNA sequence with many resulting in protein alterations and genetic malfunctions. Missense mutations occur when a change in a nucleic acid results in a change in the amino acid selected for that codon, changing the protein's structure and potentially its function. A nonsense mutation occurs when a mutation turns a codon into a stop codon, which causes the ribosome to release an incomplete protein, known as a truncated protein. Insertion mutations happen when one or more nucleotides are added to the DNA sequence. A deletion mutation is the opposite of an insertion mutation; it takes away one or more nucleotides from the DNA sequence. Both insertion and deletion mutations result in a frame shift where the ribosome is now reading incorrectly, typically resulting in a nonfunctional protein (Medline Plus). Proteins are defined as the building blocks of life, because they are essential for countless cellular processes, repairing damage, and providing structure in the body.

BRCA1/2, PTEN, and ATM genes are tumor suppressor genes that are vital in the human body, and play a significant role in cancer risk for women when mutated. 39% of women in the United States develop cancer in their lifetime. (American Cancer Society, 2025). This paper focuses on the molecular characterization of BRCA1/2, PTEN, and ATM mutations, their clinical impact, sequencing technologies for detection, and other aspects regarding women's health.

BRCA1/2 Mutations

In a normal healthy cell, BRCA1 and BRCA2 are protein coding genes which act as tumor suppressors by playing a crucial role in DNA repair (Varol et al., 2018). Cells experience DNA damage due to a number of factors, either by intercellular (e.g. replication stress) or extracellular (e.g. ultraviolet radiation damage and environmental toxins) sources (P. Mao et al., 2016). During mitosis (a process in which a cell divides and produces two identical cells), DNA strands are at a much higher risk of breaking. Additionally, overproduction of reactive oxygen

species (ROS), which are byproducts of cellular metabolism (energy production), can also cause DNA damage (Weon Yi et al., n.d.). Exposure to external factors like ultraviolet light causes DNA lesions, which can become DNA breaks. Other carcinogens, like tobacco, also induce DNA damage (Dunn et al., 2006).

BRCA1 is a highly pleiotropic DNA damage response (DDR) protein, meaning a protein that acts on multiple cellular processes in detecting and repairing DNA damage. BRCA1 plays a role in checkpoint activation, a process where a cell detects DNA damage and causes a chain reaction which pauses the cell cycle and prevents damaged DNA from being replicated during mitosis, giving time for the damaged DNA to be repaired. In addition to this, BRCA1 helps decide which repair pathway (homologous recombination or non-homologous end-rejoining) to use in cases of double-stranded DNA breaks. BRCA2 is also a pleiotropic DDR protein, because of its roles in homologous recombination and DNA replication (Roy et al., 2011).

When the cell experiences a double-strand DNA break, it decides, with the aid of BRCA1, whether to use homologous recombination or nonhomologous end rejoining. (Roy et al., 2011). In homologous recombination, BRCA1 recognizes the double-stranded DNA break and performs end resection to produce ssDNA by degrading the 5' ends of the DNA break into 3' single-stranded DNA tails. Then, RPA (a single-stranded DNA-binding protein complex) binds and stabilizes the ssDNA, preventing it from being degraded further. Consequently, BRCA2 aids in the process of replacing the RPA with the protein RAD51. RAD51 forms a nucleoprotein filament that binds an identical undamaged homologous DNA molecule to the DNA break. (Alberts B et al., 2002). Once the binding is complete, DNA polymerase extends the ssDNA and uses the undamaged homologous copy to correct the broken DNA, restoring the structure of the once-broken DNA (Liu & Huang, 2016).

In nonhomologous end-rejoining, the double stranded breaks are directly ligated together, without a homologous template (Prakash, Rohit et.al, 2015). A Ku70/Ku80 heterodimer (a DNA-binding protein complex made of the proteins Ku70 and Ku80) binds to DNA-PKcs (a protein kinase catalytic subunit) to form the DNA-PK holoenzyme, allowing the DNA ends to be aligned for end-processing. In end processing, enzymes modify the broken DNA ends so they can be bonded together. During this process, nucleotides are often added or removed, which can introduce errors (small insertions or deletions) at the repair site (Stinson & Loparo, 2022). After end-processing, enzymes and polymerases modify the DNA ends to make them ligatable (able to be processed by DNA ligase). Finally, the DNA Ligase IV allows the DNA ends to be joined again (Philippe et al., 2025).

Homologous recombination is a slower and more accurate process, because it uses extensive processing of the DNA break, formation of the nucleoprotein filament, and locating the homologous DNA pair. On the contrary, non-homologous end re-joining is a faster process but is prone to error because it lacks a homologous template to guide the repair (Z. Mao et al., 2008). Homologous recombination highly relies on BRCA1/2. If BRCA1/2 is mutated non-functionally, homologous recombination can become impaired, causing the cell to rely more on non-homologous end rejoining to repair double-strand DNA breaks. Heavy reliance on

non-homologous end-rejoining can lead to mutations and genomic instability (Prakash, Rohit et.al, 2015).

For any gene, including BRCA1/2, a person inherits two copies, one from each of their parents. An inherited BRCA1/2 gene mutation gives you a nonfunctional allele, rendering one copy of the gene unfunctional (Medline Plus). Inherited BRCA mutations cause much higher susceptibility to cancer, because unlike people with two healthy copies of the BRCA genes, they only need one somatic mutation to lose BRCA functions in their cells (Varol et al., 2018). When the second copy is mutated nonfunctionally, the cell has no functional BRCA genes left (Varol et al., 2018). This can happen through a somatic mutation or other damage to the remaining copy, like UV damage, which is caused when radiation causes thymine dimers to cause DNA breakage (P. Mao et al., 2016). When BRCA genes stop working, DNA damage may proceed unchecked, which triggers the obstruction of cell apoptosis (programmed cell death to eliminate dysfunctioning cells) (Gorodetska et al., 2019). Without the ability to apoptose or properly check DNA damage before replicating, DNA mutations can expand and accumulate. If these mutations are present in genes responsible for cellular proliferation (cell division), a cell can grow uncontrollably, a hallmark feature of cancer (Hanahan & Weinberg, 2000). As mentioned before, it's this failure to repair before cell replication that allows mutations to spread across the genome to other critical genes to propagate, like TP53, PTEN, and RB1. These genes are tumor suppressor genes like BRCA1/2, also aiding in regulating the cell cycle. When BRCA1/2 mutations spread to them, they can no longer function as they are supposed to. All these factors allow cell growth to continue unrestrained.

BRCA1 mutations often cluster on exon 11, while BRCA2 mutations often cluster on exons 10, 11, and 27 (Karami & Mehdipour, 2013). BRCA1 mutations are often small deletions/insertions that cause frameshifts and point mutations (missense and nonsense mutations), which result in truncated proteins (Karami & Mehdipour, 2013)(Medline Plus). BRCA2 mutations are commonly deletion mutations that cause frameshifts, rarely other mutation types, which cause nonfunctional proteins (Karami & Mehdipour, 2013)(Medline Plus). While normal BRCA genes make full-length proteins that function in tumor suppression and DNA repair, mutated BRCA genes are often short and unable to function properly as a tumor suppressor. They can't work with DNA repair proteins, like RAD51, anymore. (Varol et al., 2018).

Mutations in BRCA1 and BRCA2 genes significantly increase the risk of developing various cancers, especially breast and ovarian cancers in women. A comprehensive meta-analysis study of 1,948 American families, using the Cancer Genetics Network, estimated that BRCA1 mutation holders have a 46% risk of developing breast cancer and a 39% risk of developing ovarian cancer by age 70. Additionally, the same meta-analysis suggests that BRCA2 mutation holders have a 43% risk of developing breast cancer and a 22% risk of developing ovarian cancer by age 70 (Chen et al. 2006). In contrast, only about 12% of women in the US will develop breast cancer and 1.37% will develop ovarian cancer in their lifetime (National Breast Cancer Foundation, 2019)(Pearce et al., 2016).

PTEN Mutations

Regularly, the PTEN gene allows the cell to make proteins that produce a phosphatase enzyme that acts as a tumor suppressor and aids in regulating cell mitosis and keeping cells from unrestrained proliferation (Georgescu, 2010). The PTEN enzyme regulates a chemical pathway, called the PI3K/AKT pathway, that is responsible for cell proliferation. In the PI3K/AKT pathway, PI3K enzymes add phosphate groups to the lipid PIP2 to make a molecule called PIP3, which stays in the inner layer of the plasma membrane. PIP3 finds proteins with a pleckstrin homology (PH) domain, like Akt. Because of its PH domain, Akt can bind to PIP3, causing Akt to change shape, allowing for Akt phosphorylation and activation. Akt's activity promotes cell proliferation and survival (Vara, et. al, 2003). Unlike AKT, PTEN is a lipid phosphate, meaning it can remove phosphate groups from the molecule PIP3. PTEN acts as a negative regulator when it turns PIP3 into PIP2 to the PI3K/AKT pathway and restrains excess Akt activation. This is important for regulating cell reproduction, by preventing cells from uncontrolled proliferation and in performing cell apoptosis (cell death) when needed. PTEN is a critical regulator of the PI3K/AKT pathway. PTEN regulates the pathway and to make sure PIP3 only assembles when needed, when external signals present themselves (Georgescu, 2010). When a person experiences cellular damage, it can be sorted into three main situations: high stimulus, medium to low stimulus, and homeostatic state. A high stimulus state would occur in scenarios like injury or in development and tissue regeneration. In a high stimulus state, PTEN is inhibited, causing PIP3 levels rise and Akt is activated, allowing for cells to migrate to the wound and increase cell proliferation. A medium to low stimulus would occur in situations with low nutrients; PTEN dephosphorylates PIP3 back to PIP2 to conserve energy by reducing PI3K/AKT pathway activity. In a homeostatic state, the cell is balanced without injury or lack of nutrients. In this state, PI3K/AKT pathway activity is maintained but restricted to basal levels by PTEN, since the cell is at homeostasis (Cao et al., 2011).

The PTEN enzyme also aids in a range of cellular processes other than controlling excess cell growth, such as controlling cell migration, adhesion of cells to nearby tissues, creation of new blood vessels, and stabilizing genetic information (Medline Plus). Like all genes, humans receive two copies of PTEN, one from each of their parents. Losing or having one mutated PTEN gene is enough to cause inhibited cellular function and cancer (Li et al., 1997). People with hereditary PTEN gene mutations have an increased risk of developing various cancers (Luongo et al., 2019).

PTEN mutations observed in the clinic often happen in its phosphate domain, stopping PTEN from regulating PIP3 and dephosphorylating it back into PIP2 (Vara, et. al, 2003). Without a functional PTEN gene, PIP3 molecules accumulate causing the PI3K/AKT pathway to continue unrestrained and constantly activated, leading to over-production and unrestrained growth of cells without a way to apoptose. Additionally to this, PTEN does not have functional redundancy (other proteins from itself that can perform the same role in the cell), unlike PI3K and PKB/AKT in the PI3K/AKT pathway. Other proteins, like SHIP proteins, can only temporarily or subtly regulate and change PIP3 levels; PTEN is the main regulator (Chow &

Baker, 2006). Mutations can also cause structural changes in the protein, leading to irregular, short proteins that don't have functional domains, decreasing protein stability. The lack of functional enzymes means that PTEN can't act as a tumor suppressor anymore, causing misregulation of cell reproduction and contributing to cancer. PTEN is incredibly important and a mutated PTEN gene becomes very dangerous (Chow & Baker, 2006).

Inheriting or developing a PTEN gene mutation increases the risk of developing many different cancers, especially those in women. The PTEN gene is considered to be a susceptibility gene for Cowden syndrome, a disease associated with high rates of breast, thyroid, and endometrial cancer (Waite & Eng, 2002). Women with inherited PTEN mutations are 13.5 times more at risk for endometrial cancer, 8 times more at risk for breast cancer, and about 3.77 times more at risk for ovarian cancer (Cummings et al., 2023). A study on PTEN mutations in endometrial cancers found that PTEN is the most frequently mutated gene in endometrial cancer, with PTEN mutations being present in about 43% of tumors in the study (Risinger et al., 1997).

ATM Mutations

The ATM gene is a protein-coding gene on chromosome 11 that encodes for a serine/threonine kinase (STK) enzyme (Ueno et al., 2022)(Roberts & Klein, 2013). ATM is a PIKK (phosphatidylinositol 3-kinase-related kinase) gene. The ATM protein has a kinase domain, which shares the same structure as PI3K. The ATM protein has two functional domains, the FRAP-ATM-TRRAP (FAT) domain and a FRAP-ATM-TRRAP C-terminal (FATC) domain, which allow the kinase enzymes to move phosphate groups from ATP (source of cellular energy) to another molecule to release the energy from the ATP (Ueno et al., 2022). The FAT domain is a large region that wraps around the kinase domain to regulate and stabilize the ATM protein. The FATC domain is the short end of the ATM gene. Like the FAT domain, the FATC domain also stabilizes the kinase domain. In addition to this, the FATC domain works with the FAT domain and the PIKK domain to regulate the kinase active site (Jiang et al., 2006). Without FATC, ATM would not be able to use its kinase domain, consequently ATM would not be able to perform DNA damage repair (Milanovic et al., 2021).

The ATM protein has an N-terminal region, which holds many α -helical HEAT repeat motifs that are imperative for interacting with other proteins and DNA. The HEAT repeat motifs allow ATM to find the damaged DNA and bind to other regulatory and repairing proteins to facilitate repair (Phan et al., 2021). Because ATM is an STK enzyme, it is critical in repairing DNA damage and regulating ROS (reactive oxygen species) levels in cells (Ueno et al., 2022). When the ATM gene phosphorylates, it changes the protein. Because of this, ATM is able to control signaling pathways that contribute to DNA damage repair and regulation of ROS. Excess ROS can induce DNA damage, emphasizing the importance of the ATM gene in preserving the genome (Ditch & Paull, 2012).

ATM uses signaling pathways to aid various other functions in the cell, such as, apoptosis, metabolism, proliferation, and stopping the cell cycle when needed. Because of these roles, ATM is thought to be a major tumor suppressor (Ditch & Paull, 2012).

ATM mutations are linked to tumorigenesis, the process of normal cells transforming into cancer cells by the formation of tumors, across various cancer types. Normally, the ATM gene is able to pause the cell cycle, start DNA repair, and even initiate cell apoptosis when needed. When an ATM gene is mutated, the cell cannot effectively detect or repair double-stranded DNA breaks, adequately perform checkpoint activation, or signal apoptosis (Choi et al., 2016). This uncontrolled cell division causes mutations to accumulate and cause genomic instability in the cell (Milanovic et al., 2021). Without its ability to stop the cell division of damaged cells, cancer cells can survive and continue replicating (Choi et al., 2016). Additionally, there is evidence that ATM-deficient cells become resistant to cancer treatments like radiation and chemotherapies (Lee, 2024). Conversely, in late-stage tumor cells ATM can be overexpressed and further the survival of cancer cells and resistance to treatments like radiation and chemotherapy (Choi et al., 2016).

Studies suggest that around 85% of ATM mutations are truncating mutations, which result in shortened, non-functioning proteins and even proteins that will be prematurely terminated. As a result, the cell cannot properly perform DNA repair, cell cycle regulation, or apoptosis (Lavin et al., 2004). Cases have also been observed where a missense mutation can abolish ATM's kinase activity. Regularly, ATM functions in pairs of proteins, but when a mutant protein pairs with a functional protein, the bond does not work nor can it perform its actions correctly. Missense mutations account for about 15% of ATM mutations and cause the loss of protein kinase activity, meaning the protein's ability to phosphorylate. This is extremely significant; not only does the cell have to manage with nonfunctional ATM proteins, its working proteins are also being blocked by the nonfunctional ATM proteins (Xin et. al, 2019).

ATM mutations play a big role in the development of several cancers, such as breast cancer, lung cancer, thyroid cancer, and more (Choi et. al, 2016). Women with ATM gene mutations have a 25% chance of developing breast cancer and a 5% chance of developing ovarian cancer, a substantial difference from the national average of 12.5% and 1.1% respectively (Seca et. al, 2024).

Sequencing and Treatment

BRCA ½ gene mutations affect an estimation of one in four hundred people in the United States, or about 0.2-0.3% of the population (Cleveland Clinic). However, experts believe that one in eight women will develop breast cancer in their lifetime (National Breast Cancer Foundation). Women who are BRCA1/2 mutation carriers have an increased risk of various cancers, most prominently breast cancer, with a 60% chance of developing breast cancer in their lifetime (National Cancer Institute). If a woman carries a PTEN mutation, she experiences various increased risks for cancer in her lifetime, including 41-60% more likely to develop breast cancer, 28% more likely to develop endometrial cancer, and 35% more to develop thyroid cancer (Facing Hereditary Cancer Empowered). A woman with an ATM mutation has an approximately 25% increased risk of developing breast cancer in her lifetime.

Although these mutations are rare, they are highly significant because of their significant impact on women's cancer risk. Genetic testing can be used to identify these mutations, but they are not that prevalent. In 2019, a study in Georgia showed that only 24.1% of women with diagnosed breast cancer and 30.9% with diagnosed ovarian cancer had taken any genetic testing to find pathogenic gene mutations (Kurian et al., 2019). Genetic testing works by finding specific changes in a person's genes (mutations). For inherited gene mutations, a doctor or healthcare provider reviews the patient's family medical history and recommends genetics testing from there. The test is done by taking a sample of bodily fluid or tissue, like a person's blood or saliva. The objective of this is to identify cancer susceptibility and engage in preventative measures (National Cancer Institute). For somatic mutations (tumors), testing is performed on tumor tissue or leukemia/lymphoma samples. The objective of testing a somatic mutation is to find the mutation that drives the cancer development and target possible treatments (National Cancer Institute).

When a sample is taken, for either an inherited or somatic mutation screen, the genetic testing laboratory analyzes the sample, using NGS (Next-generation sequencing) technology for gene panel testing. Gene panel testing is when the lab chooses a group of genes that are known to increase cancer risk and uses sequencing technology to test for mutations. Using the sample, DNA is extracted and cut into many (millions) fragments and fixed onto a surface. Then, sequencing reactions can be performed. A detector reads the light signals that nucleotides release in the DNA fragments, enabling them to determine the DNA sequence of each fragment. Using technology, the fragments are aligned in parallel and read for mutations. This process makes it possible to examine many different genes at once (Satam et al., 2023). The panels screen for high-penetrance genes (gene mutations that between 70-100% increased cancer risk) like BRCA1/2, PTEN, TP53, MLH1, MSH2, STK11, and APC; they also screen for moderate-penetrance genes (gene mutations that cause between 30-60%) like ATM, PALB2, and CHEK2 (Angeli et al., 2020).

Overall, the objective of these genetic screenings is to improve the health of patients before or after developing cancer. Treatment varies depending on the mutations identified. For instance, some mutations may allow for targeted therapies, while others are more dependent on lifestyle changes and preventative measures.

For a BRCA 1/2 gene mutation, people can opt for a preventative treatment, such as a bilateral mastectomy, or an eliminative treatment, like PARP inhibitors. In a long-term study, 26 women with BRCA1/2 mutations who had a prophylactic (preventative) bilateral mastectomy (surgical removal of both breasts) were examined. After 13.4 years, the patients underwent check-ups and none of the women had developed breast cancer. The study estimates that prophylactic bilateral mastectomies reduce the risk of developing breast cancer by 89.5-100% (Hartmann et al., 2001). Regularly, the PARP enzyme aids in the repair of single-stranded DNA breaks. In BRCA1/2 mutated cells, PARP allows DNA damage to accumulate. PARP inhibitors block the PARP enzyme (Helleday, 2011). PARP inhibitors (a form of chemotherapy), like olaparib, have significant treatment benefits for patients who have developed BRCA1/2-mutated

cancers. A clinical study comparing a PARP inhibitor to a placebo found that olaparib treatment in patients with BRCA-mutated ovarian cancer resulted in a 70% decreased risk of disease advancement and death compared to the placebo pill (Moore et al., 2018). For PTEN, hysterectomy is a possible option, with no specific targeted treatments for it yet. A prophylactic hysterectomy can be performed on a patient with a PTEN mutation, as a preventative measure against cancer development. A hysterectomy is a surgical procedure in which a woman's uterus is removed (Facing Hereditary Cancer Empowered). ATM mutations, being moderate-penetrance genes, have less preventative options (Genomics Education). Surgery is a possibility only in high-risk individuals, while PARP treatments are being studied (Myriad Genetics, 2023).

However, genetic testing and treatment for harmful gene mutations can be costly and inaccessible for many people, with the price of getting a genetic test ranging from \$300 to \$10,000 (McDowell, 2019). This especially high cost makes testing inaccessible and poses an issue for people without health insurance. Insurance coverage for these genetic tests is variable. Many insurers don't offer coverage for these tests, because they do not deem the examinations as "medically necessary" (Smith et al., 2021). Preventative genetic testing for gene mutations before the development of cancer are often not covered by insurers, making it difficult for people to plan early intervention against cancer (National Human Genome Institute). Even if insurers do cover genetic testing, they often only cover a fraction of the cost, forcing patients to pay large out-of-pocket expenses (American Cancer Society). Insurance companies often ask for medical records and family history when making a plan for a buyer; if someone has a cancer-causing genetic mutation, this can cause their rate to increase or have them dropped altogether, because insurance companies view them as a greater financial-risk. However, the Genetic Information Nondiscrimination Act (GINA), passed in 2008, has brought significant progress to making genetic testing more accessible by reducing fears of insurance-based discrimination. GINA prohibits health insurers from using a person's genetic information to determine eligibility and coverage. GINA has several limitations, such as not protecting individuals from discrimination when applying for life or disability insurance and not requiring insurance companies to cover the cost of cancer-related genetic testing (American Cancer Society). Although awareness and availability of genetic testing continues to expand, policy reform and insurance coverage expansion needs to continue to increase accessibility and reduce disparity.

CONCLUSION

This paper has explored BRCA1/2, PTEN, and ATM gene mutations in regards to the role they play in women's cancer. Understanding cancer at a cellular level is crucial in regards to developing treatment and preventative measures. This paper has provided an overview of how cancer begins at a cellular level and how it continues to progress uncontrolledly. It touches on technological breakthroughs in cancer detection, treatment options for female specific cancers, and the obstacles of modern healthcare in getting proper diagnosis and treatment. Overall, this paper highlights the necessity of cancer screening and research on gene mutations to maintain and improve women's healthcare.

Works Cited

- Alberts, Bruce, et al. *Molecular Biology of the Cell*. 4th Edition. Garland Science, 2002, <https://www.ncbi.nlm.nih.gov/books/NBK26829/>.
- American Cancer Society. "Lifetime Risk of Developing or Dying From Cancer." American Cancer Society, 30 2025, <https://www.cancer.org/cancer/risk-prevention/understanding-cancer-risk/lifetime-probability-of-developing-or-dying-from-cancer.html>.
- Baba, Alecsandru Ioan, and Cornel Cătoi. *Comparative Oncology*. The Publishing House of the Romanian Academy, 2007, <https://www.ncbi.nlm.nih.gov/books/NBK9553/>.
- Bose, Smith, et al. *Glucose Metabolism in Cancer: The Warburg Effect and Beyond*. May 2001, <https://www.ncbi.nlm.nih.gov/books/NBK573693>.
- Brown, Joel S, et al. *Updating the Definition of Cancer*. Nov. 2023, <https://aacrjournals.org/mcr/article/21/11/1142/729785>.
- Deng, Yibin, et al. *Telomere Dysfunction and Tumor Suppression-the Senescence Connection*. June 2013, <https://pmc.ncbi.nlm.nih.gov/articles/PMC3688269/>.
- "Endometrial Cancer Prevention." *Facing Hereditary Cancer Empowered*, <https://www.facingourrisk.org/info/risk-management-and-treatment/screening-and-risk-reduction/by-cancer-type/endometrial/prevention>.
- Gambhir, Sanjiv Sam. *Molecular Imaging of Cancer with Positron Emission Tomography*. Sept. 2002, <https://www.nature.com/articles/nrc882>.
- Brown, Terence A. *Genomes*. 2nd Edition. 2002, <https://www.ncbi.nlm.nih.gov/books/NBK21114/>.
- "Gene Mutations in Humans." Cleveland Clinic, <https://my.clevelandclinic.org/health/body/23095-genetic-mutations-in-humans>.
- Liberti, Maria V, and Jason W Locasale. *The Warburg Effect: How Does It Benefit Cancer Cells?* Mar. 2017, <https://pmc.ncbi.nlm.nih.gov/articles/PMC4783224/>.
- Lugauer, Felix, and Jens Wetzl. *Medical Imaging Systems: An Introductory Guide*. Springer, 2018, <https://www.ncbi.nlm.nih.gov/books/NBK546152>.
- Philip, Joseph. *Magnetic Resonance Imaging An Advanced Modality for NonInvasive Medical Imaging*. July 2023, <https://www.openaccessjournals.com/articles/magnetic-resonance-imaging-an-advanced-modality-for-noninvasive-medical-imaging.pdf>
- Vander Heiden, Matthew G, et al. *Understanding the Warburg Effect: The Metabolic Requirements of Cell Proliferation*. May 2009, <https://www.science.org/doi/10.1126/science.1160809>.
- "What Is a Gene Variant and How Do Variants Occur?" Medline Plus, <https://medlineplus.gov/genetics/understanding/mutationsanddisorders/genemutation/>.
- "What Should I Know Before Getting Genetic Testing?" American Cancer Society, <https://www.cancer.org/cancer/risk-prevention/genetics/genetic-testing-for-cancer-risk/should-i-get-genetic-testing-for-cancer-risk.html>.

- Smith, Hadley Stevens, et al. "Outcomes of Prior Authorization Requests for Genetic Testing in Outpatient Pediatric Genetics Clinics." *Genetics in Medicine*, May 2021, <https://www.gimjournal.org/article/S1098-3600%2821%2901452-0/fulltext>.
- "Coverage and Reimbursement of Genetic Tests." National Human Genome Research Institute, 6 Feb. 2024, <https://www.genome.gov/about-genomics/policy-issues/Coverage-Reimbursement-of-Genetic-Tests>.
- McDowell, Sandy. "Cancer Patients May Not Be Told About Costs of Genomic Testing." American Cancer Society, 1 Nov. 2019, <https://www.cancer.org/research/acs-research-news/cancer-patients-may-not-be-told-about-costs-of-genomic-testing.html>.
- Alberts B, et al. "Molecular Biology of the Cell. 4th Edition." National Library of Medicine, 2002, <https://www.ncbi.nlm.nih.gov/books/NBK26898/>.
- "BRCA: The Breast Cancer Gene." National Breast Cancer Foundation, 1 Apr. 2025, <https://www.nationalbreastcancer.org/what-is-brca/>.
- Chen, Sining. Characterization of BRCA1 and BRCA2 Mutations in a Large United States Sample. Feb. 2006, <https://pubmed.ncbi.nlm.nih.gov/16484695/>.
- Dunn, Jessica, et al. "Activation of the Fanconi Anemia/BRCA Pathway and Recombination Repair in the Cellular Response to Solar Ultraviolet Light." *AACR Journals*, 2006, <https://aacrjournals.org/cancerres/article/66/23/11140/526598/Activation-of-the-Fanconi-Anemia-BRCA-Pathway-and>.
- Gorodetska, Lelizaveta, et al. "BRCA Genes: The Role in Genome Stability, Cancer Stemness and Therapy Resistance." PubMed Central, May 2019, <https://pmc.ncbi.nlm.nih.gov/articles/PMC6548160/>.
- Hanahan, Douglas, and Robert A Weinberg. *The Hallmarks of Cancer*. Jan. 2000, [https://www.cell.com/cell/fulltext/S0092-8674\(00\)81683-9?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0092867400816839%3Fshowall%3Dtrue](https://www.cell.com/cell/fulltext/S0092-8674(00)81683-9?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0092867400816839%3Fshowall%3Dtrue).
- Karami, Fatemeh, and Parvin Mehdipour. "A Comprehensive Focus on Global Spectrum of BRCA1 and BRCA2 Mutations in Breast Cancer." National Library of Medicine, Nov. 2013, <https://pmc.ncbi.nlm.nih.gov/articles/PMC3838820>.
- Liu, Ting, and Jun Huang. *DNA End Resection: Facts and Mechanisms*. May 2016, <https://pmc.ncbi.nlm.nih.gov/articles/PMC4936662>.
- Mao, Peng, et al. "UV-Induced DNA Damage and Mutagenesis in Chromatin." National Library of Medicine, Nov. 2016, <https://pmc.ncbi.nlm.nih.gov/articles/PMC5315636>.
- Mao, Zhiyong, et al. Comparison of Nonhomologous End Joining and Homologous Recombination in Human Cells. Aug. 2008, <https://pubmed.ncbi.nlm.nih.gov/18675941/>.
- Pearce, Celeste Leigh, et al. Population Distribution of Lifetime Risk of Ovarian Cancer in the United States. June 2016, <https://pmc.ncbi.nlm.nih.gov/articles/PMC4892114/>.

- Philippe, Frit, et al. "Structural and Functional Insights into the Interaction between Ku70/80 and Pol X Family Polymerases in NHEJ." 06 May 2025, May 2025, <https://doi.org/10.1038/s41467-025-59133-2>.
- Rohit Prakash, Yu Zhang, Weiran Feng, Maria Jasin. Homologous Recombination and Human Health: The Roles of BRCA1, BRCA2, and Associated Proteins. Apr. 2015, <https://pmc.ncbi.nlm.nih.gov/articles/PMC4382744/>.
- Roy, Rohini, et al. "BRCA1 and BRCA2: Different Roles in a Common Pathway of Genome Protection." *Nature Reviews Cancer*, Dec. 2011, <https://www.nature.com/articles/nrc3181>.
- Stinson, Benjamin M, and Joseph J Loparo. Repair of DNA Double-Strand Breaks by the Non-Homologous End Joining Pathway. June 2022, <https://pmc.ncbi.nlm.nih.gov/articles/PMC8899865>.
- Varol, Umut, et al. "BRCA Genes: BRCA 1 and BRCA 2." *Journal of BUON*, 2018, <https://jbuon.com/archive/23-4-862.pdf>.
- Weon Yi, Yong, et al. "BRCA1 and Oxidative Stress." *PubMed Central*, 4-3-2014, <https://pmc.ncbi.nlm.nih.gov/articles/PMC4074803/>.
- "What Is a Gene Variant and How Do Variants Occur?" *Medline Plus*, <https://medlineplus.gov/genetics/understanding/mutationsanddisorders/genemutation/>.
- "What Is a Gene Variant and How Do Variants Occur?" *Medline Plus*, <https://medlineplus.gov/genetics/understanding/mutationsanddisorders/genemutation/>.
- Cao, Lin, et al. Downregulation of PTEN at Corneal Wound Sites Accelerates Wound Healing through Increased Cell Migration. Apr. 2011, <https://pmc.ncbi.nlm.nih.gov/articles/PMC3080732>.
- Chow, Lionel M.L., and Suzanne J. Baker. PTEN Function in Normal and Neoplastic Growth. Jan. 2006, <https://www.sciencedirect.com/science/article/abs/pii/S0304383505010219?via%3Dihub>.
- Cummings, Shelly, et al. Cancer Risk Associated With PTEN Pathogenic Variants Identified Using Multigene Hereditary Cancer Panel Testing. Jan. 2023, <https://ascopubs.org/doi/10.1200/PO.22.00415>.
- Georgescu, Maria-Magdalena. "PTEN Tumor Suppressor Network in PI3K-Akt Pathway Control." *National Library of Medicine*, Dec. 2010, <https://pmc.ncbi.nlm.nih.gov/articles/PMC3092286/>.
- Li, Jing, et al. PTEN, a Putative Protein Tyrosine Phosphatase Gene Mutated in Human Brain, Breast, and Prostate Cancer. Mar. 1997, <https://www.science.org/doi/10.1126/science.275.5308.1943>.
- Luongo, Francesca, et al. PTEN Tumor-Suppressor: The Dam of Stemness in Cancer. July 2019, <https://pmc.ncbi.nlm.nih.gov/articles/PMC6721423>.
- "PTEN Gene." *Medline Plus*, <https://medlineplus.gov/genetics/gene/pten/>.

- Risinger, John I, et al. PTEN/MMAC1 Mutations in Endometrial Cancers. Nov. 1997, <https://aacrjournals.org/cancerres/article/57/21/4736/503681/PTEN-MMAC1-Mutations-in-Endometrial-Cancers>.
- Vara, Juan Ángel Fresno. PI3K/Akt Signalling Pathway and Cancer. Nov. 2003, https://www.sciencedirect.com/science/article/abs/pii/S0305737203001622?casa_token=2VO669pwXeAAAAA:O8hGPwL9AMw_t1ecLUAV-E2_nNZL1Rxs0NqtIwd-Q6h_ylFgMq2A4ZI1SetQO5bty7RMwuuW8Q.
- Waite, Kristin A, and Charis Eng. Protean PTEN: Form and Function. Mar. 2002, <https://pmc.ncbi.nlm.nih.gov/articles/PMC379112/>.
- Waite, Kristin A., and Charis Eng. Protean PTEN: Form and Function. Apr. 2002, [https://www.cell.com/ajhg/fulltext/S0002-9297\(07\)60291-2](https://www.cell.com/ajhg/fulltext/S0002-9297(07)60291-2).
- Choi, Michael, et al. “ATM Mutations in Cancer: Therapeutic Implications.” American Association for Cancer Research, Aug. 2016, <https://aacrjournals.org/mct/article/15/8/1781/146010/ATM-Mutations-in-Cancer-Therapeutic>.
- Ditch, Scott, and Tanya T Paull. The ATM Protein Kinase and Cellular Redox Signaling: Beyond the DNA Damage Response. Nov. 2011, <https://pmc.ncbi.nlm.nih.gov/articles/PMC3259275>.
- Jiang, Xiaofeng, et al. The FATC Domains of PIKK Proteins Are Functionally Equivalent and Participate in the Tip60-Dependent Activation of DNA-PKcs and ATM. June 2006, <https://pubmed.ncbi.nlm.nih.gov/16603769/>.
- Lavin, Martin F, et al. Functional Consequences of Sequence Alterations in the ATM Gene. 2004, <https://pubmed.ncbi.nlm.nih.gov/15279808/>.
- Lee, Ji-Hoon. Targeting the ATM Pathway in Cancer: Opportunities, Challenges and Personalized Therapeutic Strategies. Sept. 2024, <https://www.sciencedirect.com/science/article/pii/S0305737224001361>.
- Milanovic, Maja, et al. FATC Domain Deletion Compromises ATM Protein Stability, Lymphocyte Development, and Promotes Lymphomagenesis. <https://pmc.ncbi.nlm.nih.gov/articles/PMC9305077>.
- Phan, Liem Minh, and Abdol-Hossein Rezaeian. ATM: Main Features, Signaling Pathways, and Its Diverse Roles in DNA Damage Response, Tumor Suppression, and Cancer Development. May 2021, <https://pmc.ncbi.nlm.nih.gov/articles/PMC8228802>.
- Roberts, Nicholas J, and Alison P Klein. Next Generation Sequencing Applications in Cancer Research. 2013, <https://www.sciencedirect.com/topics/medicine-and-dentistry/atm-protein>.
- Seca, Marta, and Steven A Narod. Breast Cancer and ATM Mutations: Treatment Implications. Nov. 2024, <https://pmc.ncbi.nlm.nih.gov/articles/PMC11566136/>.
- Ueno, Sayaka, et al. ATM: Functions of ATM Kinase and Its Relevance to Hereditary Tumors. Jan. 2022, <https://pmc.ncbi.nlm.nih.gov/articles/PMC8745051/>.
- Xin, Jiyu, et al. “Structural Basis of Allosteric Regulation of Tel1/ATM Kinase.” Cell Research, 2019, <https://www.nature.com/articles/s41422-019-0176-1>?

Angeli, Davide, et al. Genetic Predisposition to Breast and Ovarian Cancers: How Many and Which Genes to Test? Feb. 2020, <https://pmc.ncbi.nlm.nih.gov/articles/PMC7038038>.

“ATM Gene - Cancer Risk, Management, and Testing for Genetic Mutations.” Myriad Genetics, <https://myriad.com/gene-results/ATM/1>.

“BRCA: The Breast Cancer Gene.” National Breast Cancer Foundation, 1 Apr. 2025, <https://www.nationalbreastcancer.org/what-is-brca/>.

Efficacy of Bilateral Prophylactic Mastectomy in BRCA1 and BRCA2 Gene Mutation Carriers.

“Endometrial Cancer Prevention.” Facing Hereditary Cancer Empowered, <https://www.facingourrisk.org/info/risk-management-and-treatment/screening-and-risk-reduction/by-cancer-type/endometrial/prevention>.

“Gene Panel Sequencing.” Genomics Education, <https://www.genomicseducation.hee.nhs.uk/genotes/knowledge-hub/gene-panel-sequencing>.

“Genetic Testing for Inherited Cancer Risk.” National Cancer Institute, <https://www.cancer.gov/about-cancer/causes-prevention/genetics/genetic-testing-fact-sheet#what-is-genetic-testing-for-inherited-cancer-risk>.

Hartmann, Lynn C., et al. Efficacy of Bilateral Prophylactic Mastectomy in BRCA1 and BRCA2 Gene Mutation Carriers. Nov. 2001, <https://mayoclinic.elsevierpure.com/en/publications/efficacy-of-bilateral-prophylactic-mastectomy-in-brca1-and-brca2->.

Helleday, Thomas. The Underlying Mechanism for the PARP and BRCA Synthetic Lethality: Clearing up the Misunderstandings. July 2011, <https://pubmed.ncbi.nlm.nih.gov/21821475/>.

Kurian, Allison W, et al. Genetic Testing and Results in a Population-Based Cohort of Breast Cancer Patients and Ovarian Cancer Patients. Apr. 2019, <https://ascopubs.org/doi/full/10.1200/JCO.18.01854>.

Moore, Kathleen, et al. Maintenance Olaparib in Patients with Newly Diagnosed Advanced Ovarian Cancer. Oct. 2018, <https://www.nejm.org/doi/full/10.1056/NEJMoa1810858>.

“PTEN Gene Mutations: Cancer Risks.” Facing Hereditary Cancer Empowered, <https://www.facingourrisk.org/info/hereditary-cancer-and-genetic-testing/hereditary-cancer-genes-and-risk/genes-by-name/pten/cancer-risk>.

Satam, Heena, et al. Next-Generation Sequencing Technology: Current Trends and Advancements. July 2023, <https://www.mdpi.com/2079-7737/12/7/997>.

Holy Causes and Hidden Motives: Power and Faith in the Thirty Years' War

Introduction By Miles Snyder

Many historians describe the Thirty Years' War as the ultimate "war of religion," where Catholics and Protestants battled for decades in one of the bloodiest conflicts in European history. Textbooks often portray the conflict as the natural result of religious hatred, pointing to dramatic moments such as the Defenestration of Prague or the speeches of Gustavus Adolphus as evidence that men fought and died because of faith alone. At first glance, this explanation seems convincing because religion shaped how rulers spoke about the war, how manifestos were written, and how soldiers and civilians understood their suffering. In a deeply religious age, it makes sense that leaders would justify themselves in religious terms. Yet when the events are studied more closely, the explanation begins to fall apart. If the war was truly only about Catholic versus Protestant, then it would be impossible to explain why Catholic France entered the war on the Protestant side in 1635 or why Protestant princes often refused to work together when it threatened their own interests. The Thirty Years' War was far more complicated than a simple clash of faiths.

Politics and religion in early modern Europe were never truly separate, and the conflict shows how the two forces worked together to drive violence. Religion was not irrelevant, but it was often the language through which political ambition was expressed. Leaders used religious justifications because it gave their actions legitimacy and inspired loyalty, but their choices were guided just as much by questions of sovereignty, territory, and dynastic survival. Frederick V of the Palatinate claimed to defend Protestant liberties, but in reality he was trying to expand his influence and elevate his family's power within the Empire. Christian IV of Denmark also claimed to be fighting for Lutheranism, but his campaigns in northern Germany reveal clear strategic interests in protecting his own borders and controlling trade routes along the Baltic. Gustavus Adolphus of Sweden, though deeply Lutheran, used religious rhetoric to justify his intervention while pursuing territorial control and international prestige. Even more revealing, France, a Catholic kingdom, eventually allied itself with Protestant powers, proving that the balance of power often mattered more than religious unity. The deeper truth is that the Thirty Years' War demonstrates the inseparability of religion and politics in early modern Europe. Faith provided the vocabulary of legitimacy, but political ambition was the engine that kept the conflict alive for three decades.

Understanding this matters for more than just historical accuracy. If the Thirty Years' War is seen only as a warning about religious fanaticism, then its real lessons are lost. The conflict shows how leaders rarely fight for one pure idea but instead combine belief with ambition in ways that make violence seem both necessary and righteous. This is not just a lesson about the seventeenth century. In conflicts across the modern world, politicians still use ideology, faith, and national identity to justify wars that are just as much about power, resources, and control. Recognizing this pattern helps reveal the deeper forces that drive violence and explains why wars are so often portrayed as struggles of principle when they are equally struggles for dominance.

The Thirty Years' War is therefore not only a key moment in European history but also a warning about how dangerous it becomes when sacred language and political ambition fuse into one powerful justification for conflict.

This essay examines how rulers fused religion and politics to legitimize violence during the Thirty Years' War. By analyzing the rhetoric and actions of Frederick V, Christian IV, Gustavus Adolphus, and Queen Christina, alongside France's intervention, this essay argues that faith and power were inseparably bound together. Religion provided the sacred language through which ambition was expressed, whether to justify expansion, to build alliances, or even to legitimize abdication. The case studies show that political motives always had to be spoken in confessional terms, proving that the war was not simply religious or political but a synthesis of both.

Early Phase and start of the thirty years war

The Thirty Years' War did not begin as a straightforward religious struggle or as a simple battle for political power. Instead, it emerged from the structural tensions within the Holy Roman Empire, where fragmented authority allowed rulers to link their political ambitions with religious identities. The Peace of Augsburg of 1555 had institutionalized this connection by making the ruler's religion the official faith of each state. As Peter H. Wilson explains in "Dynasty, Constitution, and Confession: The Role of Religion in the Thirty Years' War," this arrangement ensured that "confession became the idiom of political debate" (Wilson, 12). Any challenge to political authority could be framed as a religious threat, and any religious dispute was simultaneously a political one. The result was a system where religion and politics were not separate forces but deeply intertwined, making conflict almost inevitable when interests collided.

The fragility of The Peace of Augsburg soon became evident. Calvinists, excluded from the 1555 settlement, had by 1600 taken control of key territories, most importantly the Palatinate. The unresolved issue of church lands that had been secularized further fueled antagonism. Out of this uncertainty, rival confessional blocs emerged. In 1608 Frederick IV of the Palatinate established the Protestant Union, claiming it was to defend "the true Reformed religion and the liberties of Protestant princes" (Cited in Lockhart 205). Here, political independence was cloaked in the language of faith, since liberty meant the right to resist Habsburg interference. In response, Duke Maximilian of Bavaria formed the Catholic League in 1609, pledging to preserve "the ancient Catholic religion and the rights of the Emperor" (Lockhart 206). Both alliances presented themselves as defenders of divine order, yet each simultaneously advanced the political dominance of its leader. Before war broke out, confessional identity was already the vehicle through which political coalitions were justified.

The Bohemian revolt of 1618 showed how these dynamics could erupt into violence. In 1609 Emperor Rudolf II issued the Letter of Majesty, which guaranteed Protestants in Bohemia freedom of worship and limited royal interference. When Ferdinand II, a staunch Catholic, began restricting these rights, Protestants staged the famous Defenestration of Prague. According to one participant, they acted "for the liberty of the Gospel and the ancient rights of the Estates" (Cited

in Palmitessa 143). Catholics, by contrast, denounced the deed as both “heresy and treason” (Palmitessa 145). This dual language shows how each side interpreted the same event as simultaneously religious and political. To the rebels, defending worship was inseparable from defending constitutional rights; to the crown, rebellion against the emperor was rebellion against God.

The stakes rose dramatically in 1619 when the Bohemian Estates deposed Ferdinand and offered their crown to Frederick V of the Palatinate. Frederick accepted, declaring he did so out of “duty to God and the Reformed religion” (Cited in Lockhart 214). Yet his new crown also expanded Palatine influence deep into Central Europe. Wilson emphasizes that Frederick became “the symbol of Protestant resistance,” but also placed himself in direct political opposition to the Habsburg dynasty (Wilson 14). His appeals to the Protestant Union and to his father-in-law James I of England show that dynastic and strategic concerns underpinned his religious rhetoric. Lockhart concludes that by this point the Union was “essentially an instrument of Palatine policy” (Lockhart 215). Frederick’s own words reveal how religion sanctified what was in reality a dynastic gamble.

Ferdinand II also framed his response in equally confessional terms. After Imperial forces crushed Frederick at White Mountain in 1620, Ferdinand declared that the rebels had committed “not only high treason against their natural lord but abominable heresy against Almighty God” (Cited in Wilson 15). He confiscated rebel lands, transferred Frederick’s electoral title to Bavaria, and launched a program of Catholic restoration. This culminated in the 1629 Edict of Restitution, which ordered the return of all church property secularized since 1552. The decree announced its purpose as “the restoration of holy religion and the honor of God” (Wilson 16). In practice, it dramatically expanded imperial authority at the expense of Protestant princes. Lockhart observes that the decree’s sweeping scope “united Lutherans and Calvinists in common cause” against the emperor (Lockhart 220). Once again, what was justified as a sacred settlement functioned as a major act of political consolidation.

The Danish phase reinforced this pattern. In 1625 Christian IV of Denmark, Frederick V’s uncle, entered the war claiming to defend Protestant liberties in northern Germany. He declared his goal was “to uphold the liberty of the Gospel and protect our brethren in faith” (Cited in Lockhart 222). Yet his actions also aimed to expand Danish power along the Elbe and strengthen control of Baltic trade. His defeat by Imperial generals Tilly and Wallenstein showed the limits of confessional rhetoric when it lacked strategic depth. The episode foreshadowed Sweden’s later intervention. As Roberts shows in “The Political Objectives of Gustavus Adolphus in Germany,” Sweden justified its entry in 1630 as a mission to preserve Protestantism, but Gustavus’ deeper aim was to control the Baltic coastline (Roberts 5–6). Christian IV’s failure made clear that appeals to faith were only effective when paired with sufficient resources and dynastic calculation.

The early phases of the Thirty Years’ War demonstrate that religion and politics were never separate forces but two sides of the same coin. The Peace of Augsburg had enshrined confessional identity as the basis of political legitimacy, ensuring that rulers expressed their

ambitions in religious terms. Frederick V used the defense of Protestantism to legitimize a dynastic bid for Bohemia, while Ferdinand II invoked Catholic unity to justify confiscation and centralization. Christian IV wrapped Danish expansion in the Lutheran language, just as Gustavus later fused Protestant duty with Swedish security. As Wilson concludes, the war's theological disputes were real, but they were "subordinate to disputes over the exercise of rights" in the Empire (Wilson 18). Confession was the language through which power was claimed, challenged, and resisted. In the early decades of the conflict, every act of rebellion, conquest, or consolidation was framed as both sacred and political, proving that the Thirty Years' War cannot be reduced to one motive alone.

Sweden's Entry

When Sweden entered the Thirty Years' War in 1630 under King Gustavus Adolphus, many contemporaries believed it was to save German Protestants from destruction at the hands of the Catholic Habsburgs. Gustavus presented himself as a defender of the faith and Swedish propaganda described him as the "Lion of the North" chosen by God to protect the Protestant cause. While Gustavus was a sincere Lutheran, it is misleading to see his intervention only as a religious crusade. His actions reveal that he had precise political goals tied to Sweden's security and to his own dynastic ambition. Gustavus wanted to control parts of the Baltic coast, gain influence within the Empire, and elevate Sweden's position in Europe. To accomplish these aims, he used religion as a justification that transformed political ambition into moral duty. As Frost explains, Gustavus consistently linked Sweden's survival to control of German territories (Frost 13–14), while Wilson notes that confessional rhetoric was the natural language through which rulers expressed such ambitions (Wilson 242). In the early modern world, politics and religion were inseparable. The language of faith was not a disguise for politics but the idiom through which political violence was legitimated and made comprehensible. This entanglement of faith and politics was not confined to Gustavus' reign. His daughter, Queen Christina, later demonstrated the same dynamic in a very different form, as her abdication and conversion showed how rulers continued to present personal and dynastic choices in confessional terms.

One of Gustavus' clearest goals was to secure Sweden's position in the Baltic. Sweden was a relatively small kingdom surrounded by powerful rivals, and Gustavus knew that its independence depended on controlling access to the sea. German territories such as Stralsund, Wismar, and Pomerania were especially important since whoever held them could dominate the southern coast of the Baltic and threaten Sweden directly. Frost demonstrates that Gustavus entered Germany determined to establish "a firm territorial base" in these regions, both to prevent invasion and to strengthen Sweden's long-term influence in European politics (Frost 17). This was a dynastic project as much as a defensive one. Gustavus sought to elevate the Vasa dynasty into the ranks of Europe's great powers. Yet publicly he framed these aims in religious terms, claiming that Sweden's purpose was to protect Protestantism from annihilation. Wilson emphasizes that in the political culture of the time, rulers were expected to cast strategic goals in confessional language since legitimacy was understood through faith (Wilson 245). Gustavus'

Lutheran identity thus provided a ready framework for describing territorial expansion as a moral duty.

His political ambition was evident in his 1630 treaty with Duke Bogislaw XIV of Pomerania. The treaty was presented as a Protestant alliance, but it placed Pomerania's fortresses under Swedish control and allowed Gustavus to retain the duchy after Bogislaw's death unless Brandenburg, the heir, accepted Swedish terms. Frost notes that this arrangement was part of Gustavus' larger design to build "a system of client states" in northern Germany loyal to Sweden (Frost 18). Such a system would extend Swedish influence far beyond its natural borders. Religion played an essential role here. By presenting himself as protector of Protestant liberties, Gustavus could justify garrisons and interventions that otherwise looked like foreign domination. Evans underlines that this was not empty propaganda but the framework through which order and legitimacy were conceived (Evans 245–247). Political ambition was inseparable from religious duty and the language of confessional solidarity was the medium through which dynastic expansion could be accepted.

The Swedish Manifesto of July 1630 illustrates this fusion of political ambition and religious obligation. The document portrayed Sweden's campaign as defensive, aimed at preserving Protestantism, restoring the Empire to its 1618 balance, and securing the liberties of the estates. Evans notes that the manifesto carefully avoided the language of conquest, despite Gustavus' real territorial aims (Evans 246). This rhetorical choice was deliberate. Gustavus needed allies in Germany, subsidies from England and the Dutch, and support from his own subjects. A political argument about securing ports would not suffice to mobilize men and money. By casting the war as a moral obligation, Gustavus transformed strategic ambition into a sacred cause. Frost stresses that Gustavus distinguished his aims from those of Christian IV of Denmark, whose intervention in the 1620s had been dismissed as self-serving (Frost 21). Gustavus understood that only by presenting his mission as both holy and necessary could he build a broad coalition. His manifesto exemplifies how, in the Thirty Years' War, political goals were pursued through religious rhetoric, not as deception but as part of the cultural grammar of legitimacy and violence.

Religion also shaped Gustavus' ability to mobilize resources and allies. Sweden's population was small, and sustaining a major war in Germany required heavy taxation and recruitment. A purely political argument would not have inspired such sacrifice. By framing the conflict as a cosmic struggle between God and the Devil, Gustavus created urgency and moral obligation. In July 1630 he told a Brandenburg envoy that "there is no third way," meaning that neutrality was impossible (Piirimäe 53). This absolutist rhetoric pressured German Protestant princes to choose sides and encouraged Swedish subjects to endure the burdens of war. Wilson points out that such rhetoric was typical of early modern rulers, who used confessional language to justify violence and demand loyalty (Wilson 248). Gustavus' religious framing allowed him to present setbacks as part of God's providence, sustaining morale even in adversity. At the same time, his acceptance of subsidies from Catholic France reveals his pragmatism. Religion structured the justification of war, but political necessity shaped its practice. As Frost notes,

Gustavus was a master at balancing idealism with realism, maintaining confessional rhetoric while pursuing dynastic advantage (Frost 23).

One way to see the longer legacy of Gustavus' political-religious strategy is in how Swedish rulers after him continued to frame dynastic and strategic decisions in confessional terms. Gustavus had shown that Sweden's influence in Europe depended on presenting political necessity as religious duty, and this lesson endured after his death. His chancellor Oxenstierna reminded the council in 1636 that "it was thus a question not so much of religion, as of status publicus, in which religion is comprehended" (Roberts 22). This statement reveals the fusion that shaped Sweden's policy. Political survival and confessional identity were not separate categories but parts of one framework. Religion gave political action its legitimacy, while political security safeguarded the faith. By grounding strategy in this combined language, Gustavus and his successors ensured that Swedish expansion could be narrated as both dynastic ambition and sacred obligation. It was within this cultural and political inheritance that his daughter, Queen Christina, came to power, and her reign demonstrates how deeply intertwined religion and politics remained in seventeenth-century Sweden.

Queen Christina's reign highlights the same inseparable bond between religion and politics but in a way that differed sharply from Gustavus or Denmark's rulers. Christina, who succeeded Gustavus Adolphus as his only surviving child in 1632 and ruled Sweden until her abdication in 1654, became one of the most unconventional monarchs of seventeenth-century Europe. Her abdication revealed how personal and dynastic decisions had to be expressed in religious terms to gain legitimacy. She declared, "I have sacrificed my crown and my country, which were dear to me, only for the glory of God" (as cited in Roberts 24). In this statement she transformed what might have been seen as abandonment of her realm into a sacred act of service. The rhetoric shows that even the rejection of political power required confessional language to be understood and accepted.

Christina's later conversion to Catholicism reinforced the extent to which religion shaped political meaning. By framing her conversion as a divine obligation she turned a decision that weakened Sweden's Protestant identity into an act of obedience to God. In contrast to Gustavus, who had used Lutheran language to justify Swedish expansion, or Denmark, which had cloaked dynastic ambition in Protestant rhetoric, Christina presented political withdrawal as a religious sacrifice. Her case is important for understanding the Thirty Years' War because it proves that rulers could not separate political choices from confessional frameworks. She demonstrates that religion and politics were so deeply entangled that even leaving the throne had to be justified in sacred terms. By highlighting this inversion Christina becomes essential to the larger argument of the essay, showing that faith provided legitimacy not only for conquest and expansion but also for renunciation and retreat.

The cases of Gustavus Adolphus and Queen Christina together show how religion and politics were inseparably connected in the Thirty Years' War and its aftermath. Gustavus was a sincere Lutheran, but he was also a dynastic monarch who pursued territorial security and European influence. His use of religious rhetoric was not cynical hypocrisy but the necessary

language of legitimacy in an age when confessional identity shaped political order. Christina, too, acted within this framework. She cast her refusal to marry and her abdication as matters of spiritual liberty and divine service, transforming personal and political choices into sacred obligations. Compared to Denmark's earlier intervention, Gustavus' policies were more ambitious and more carefully wrapped in moral language, which enabled him to build a broader coalition. Compared to France's later entry, Gustavus' rhetoric was still overtly confessional, reflecting an earlier phase of the war when religion and politics were spoken of as a unity. Christina's conversion shows how the same language of legitimacy could work in reverse, making the renunciation of power itself a religious duty. Together, father and daughter remind us that in seventeenth-century Europe rulers mobilized faith and dynasty together, binding sacred obligation and political necessity into a single cause.

France and the Decline of Religious Centrality

When Catholic France entered the Thirty Years' War on the side of Protestant powers in 1635, the character of the conflict changed profoundly. Up to that point, the war had largely been justified and narrated in terms of confessional identity. The Bohemian revolt in 1618 was cast by its leaders as a defense of Protestant liberties within the Empire, and Denmark's intervention in the late 1620s was framed as a Lutheran crusade against Habsburg Catholicism. Even Gustavus Adolphus of Sweden, whose strategic aims were far-reaching, presented his campaign as a mission entrusted to him by God to save German Protestantism from destruction. France's entry disrupted this pattern. Cardinal Richelieu and his advisers did not present their decision as a matter of confessional solidarity but instead explained it as a struggle for the balance of power in Europe. As Sutherland observes, Richelieu's central concern was "the fear of encirclement by Habsburg territories and the threat of universal monarchy" (Cited in Sutherland 117). In his own words, Richelieu declared, "The safety of the state is the supreme law, and everything must yield to it" (Cited in Croxton 214). This statement shows that Richelieu framed political necessity as the ultimate priority, even while using religion as a source of legitimacy. By presenting France's intervention as a defense against Habsburg hegemony rather than a war of religion, the French crown signaled a decisive shift in the way the conflict was justified and understood.

France's fears of Habsburg power had deep roots, and this concern provided the political logic for intervention. Since the sixteenth century, French rulers had worried about being surrounded by Habsburg territories. The so-called Spanish Road, a chain of possessions linking Spain and the Spanish Netherlands through Italy and Germany, allowed Habsburg troops to move efficiently across Europe. Combined with Imperial successes in Germany in the late 1620s and early 1630s, this raised the specter of France's encirclement. Richelieu believed that if the emperor consolidated his power in Germany while Spain retained its control in Italy and the Low Countries, France would be trapped between Habsburg dominions on all sides. Wilson notes that France entered the war "not to defend German Protestantism but to prevent Habsburg dominance of central Europe" (Wilson 191). Richelieu himself argued that "to weaken the House of Austria is to labor for the repose of Christendom" (cited in Croxton 217). This remark demonstrates how

he fused political calculation with sacred duty. For him, the threat of a “universal monarchy” was not an abstract fear but a direct political danger. The longer the Habsburgs consolidated their control, the weaker France would become, regardless of religion. Yet Richelieu faced a dilemma: how to explain an alliance with Protestants when France was still a Catholic kingdom.

Cardinal Richelieu resolved this dilemma by tailoring his message to different audiences. To the outside world, France presented its war aims as secular, justified in the language of balance-of-power politics. The French declaration of war in 1635 made no appeal to the defense of religion. Instead, it described Habsburg power as a threat to the liberties of Europe. The manifesto stated, “The liberty of all Europe is at stake, and the King cannot without dishonor abandon his allies to the oppression of the House of Austria” (Croxtton 220). This phrasing avoids confessional rhetoric while making political danger sound universal. Sutherland observes that the manifesto contained “not once an appeal to the defence of religion, but only to the danger of Habsburg universal monarchy” (Sutherland 119). This was a striking departure from the manifestos of Gustavus Adolphus, which had been steeped in Lutheran rhetoric. Evans emphasizes that French leaders “abandoned explicit confessional justification in favour of the secular language of state interest” (Evans 244). In domestic politics, however, Richelieu relied on Catholic imagery and rituals to shore up legitimacy. Sermons portrayed France’s struggle as the defense of Christendom’s liberties, and Richelieu himself embodied the paradox of a cardinal leading a war alongside Protestants. Rather than separating religion and politics, Richelieu combined them in different registers depending on his audience. Abroad, France emphasized political balance; at home, it emphasized Catholic duty. This demonstrates how rulers in the Thirty Years’ War operated within the shared cultural framework of religion and politics, but manipulated that framework to fit changing circumstances.

Richelieu’s diplomacy in the years immediately before 1635 reveals how carefully he managed this balancing act. France had already intervened indirectly in the conflict by supporting Sweden after 1630, funding Gustavus Adolphus and later his successors in the Heilbronn League. It had also moved militarily to occupy Lorraine in 1633 and secure the Valtelline passes that linked Spanish and Imperial possessions. These actions made little sense in purely religious terms, since France was undermining fellow Catholic rulers. But in political terms, they were consistent: each step aimed at weakening Habsburg power. Wilson explains that France’s policy was “to check the emperor while avoiding a religious war at home” (Wilson 192). Richelieu captured this logic when he wrote, “In matters of state, one must act as if religion did not exist, though never speak as if one did not respect it” (Croxtton 218). This striking statement makes clear how he subordinated confessional difference to political necessity while still recognizing the cultural importance of religious language. For Richelieu, this meant constantly rearticulating violence in terms that would resonate with different audiences. At court and in the pulpits, the language was Catholic and moral; in treaties and manifestos, it was political and pragmatic. This duality shows that Richelieu did not abandon religion but redefined its role within political calculation.

The contrast with Gustavus Adolphus highlights the significance of France's shift. Gustavus' 1630 manifesto declared that Sweden's purpose was "to preserve the true religion, to maintain the constitution of the Empire, and to secure the liberties of the estates" (Evans 245). This formula bound together confession, imperial law, and liberty in a way that made political ambition sound religious. By contrast, France's 1635 manifesto did the opposite. It spoke only of balance, liberty, and the danger of universal monarchy. Religion was not mentioned as a justification at all. As Sutherland notes, this was the first major intervention in the war whose official rationale excluded religious solidarity (Sutherland 120). The difference does not mean that Gustavus was religious and Richelieu was political, but that each used the intersection of religion and politics in different ways. Gustavus infused his strategic aims with sacred rhetoric to appeal to Protestant allies, while Richelieu reframed religious loyalty in political terms to hold together a Catholic monarchy that fought alongside Protestants. Both demonstrate how, in the early modern world, political violence was always mediated through the language of legitimacy, which could be sacred, secular, or both at once.

The entry of France also marks a broader shift in the nature of the Thirty Years' War. The early stages were driven by confessional disputes: the Bohemian revolt, the Palatinate crisis, and Denmark's Lutheran intervention were all framed in explicitly religious terms. Even Sweden's intervention, with its clear political and dynastic goals, was justified by appeals to Protestant solidarity. France's approach subordinated religion to *raison d'état*. Evans observes that Richelieu's policy marked the moment when "confessional justification gave way to the secular idiom of state interest" (Evans 246). This did not mean that religion vanished. Catholic identity still shaped French domestic politics and Protestant identity still structured alliances. But the war's center of gravity shifted. From 1635 onwards, the conflict was less about defending or restoring religious order in the Empire and more about preventing Habsburg hegemony in Europe. Wilson argues that "from 1635 onwards, the war assumed the form of a struggle between states rather than confessions" (Wilson 192). The war's violence was still narrated in religious terms at the local level, but at the highest level of diplomacy and strategy, political calculation had become dominant.

Conclusion

The Thirty Years' War is often remembered as the most destructive of the "wars of religion," but a closer look reveals that it was never just a clash between Catholics and Protestants. Religion gave rulers the words, symbols, and authority they needed to justify their choices, but the actual direction of the conflict was driven by political ambition and dynastic rivalry. Frederick V, Christian IV, and Gustavus Adolphus all spoke about defending faith, yet their real goals included protecting sovereignty, expanding territory, and strengthening their own families' power. Catholic France's decision to fight alongside Protestant powers in 1635 proves that the balance of power was often more important than religious solidarity, since survival and influence outweighed confessional loyalty. In this way the war demonstrates that religion and politics were inseparable in early modern Europe: faith was not a disguise for ambition, but the

cultural framework through which ambition was expressed. This makes the war important not only for understanding how Europe was torn apart for three decades but also for recognizing how rulers throughout history have justified violence. If the Thirty Years' War is viewed only as a cautionary tale about religious fanaticism, then the deeper lessons are missed. The conflict shows that leaders rarely admit they are fighting for power, but instead clothe their ambitions in language that inspires loyalty and sacrifice. Whether through appeals to God, to liberty, to ideology, or to nationalism, the strategy is the same: turn political goals into moral duties so that violence appears not only necessary but righteous. This insight matters because it helps explain why wars have continued to erupt long after the seventeenth century in ways that look very different on the surface but follow the same underlying pattern. The Thirty Years' War warns us that the most destructive conflicts are not caused by religion alone or politics alone but by the dangerous combination of both, when sacred language and worldly ambition fuse together into one powerful justification for conflict.

Works Cited

- Wilson, Peter H. "Dynasty, Constitution, and Confession: The Role of Religion in the Thirty Years War." *The International History Review*, vol. 30, no. 3, 2008, pp. 473-514. Taylor & Francis, <https://doi.org/10.1080/07075332.2008.10415483>.
Dynasty constitution and conf...
- Roberts, Michael. "The Political Objectives of Gustavus Adolphus in Germany, 1630–1632." *Transactions of the Royal Historical Society*, vol. 6, 1956, pp. 19-42. Cambridge University Press, <https://doi.org/10.2307/3678885>.the-political-objectives-of-gus...
- Lockhart, Paul Douglas. "Religion and Princely Liberties: Denmark's Intervention in the Thirty Years War, 1618-1625." *The International History Review*, vol. 17, no. 1, Feb. 1995, pp. 1-22. Taylor & Francis, <https://www.jstor.org/stable/40106911>.
Lockhart-ReligionPrincelyLibert
- Palmitessa, James R. "The Prague Uprising of 1611: Property, Politics, and Catholic Renewal in the Early Years of Habsburg Rule." *Central European History*, vol. 31, no. 4, 1998, pp. 299-328. Cambridge University Press, <https://doi.org/10.1017/S0008938900017040>.
- Parrott, David. *War in History*, vol. 8, no. 2, 2001, pp. 235–37. JSTOR, <http://www.jstor.org/stable/26013862>. Accessed 21 Sept. 2025.
- Lincoln, Bruce. "Theses on Religion & Violence." *Isim Review*, vol. 15, no. 1, 2005, p. 12. Leiden University Scholarly Publications, <hdl.handle.net/1887/16990>.
- Selengut, Charles. *Sacred Fury: Understanding Religious Violence*. 3rd ed., Rowman & Littlefield, 2017.
- Croxton, Derek. "The Territorial Imperative: Military–Cultural–Political Aspects of the Thirty Years' War." *German Studies Review*, vol. 21, no. 2, 1998, pp. 201–227.
- Frost, Robert I. *The Northern Wars: War, State, and Society in Northeastern Europe, 1558–1721*. Longman, 2000.
- Evans, R. J. W. "The Thirty Years' War." In *The Coming of the Thirty Years' War*, edited by Peter H. Wilson, Routledge, 1998, pp. 243–249.
- Piirimäe, Pärtel. "Religion, Force, and the Absolutist State: Swedish Confessional Rhetoric during the Thirty Years' War." *European History Quarterly*, vol. 35, no. 1, 2005, pp. 31–57. Sage Publications, <https://doi.org/10.1177/0265691405050460>.
- Sutherland, N. M. *Richelieu and Reason of State*. Cambridge University Press, 1985.
- Piirimäe, Pärtel. "Just War in Theory and Practice: The Legitimation of Swedish Intervention in the Thirty Years War." *The Historical Journal*, vol. 45, no. 3, 2002, pp. 499–523. JSTOR, <http://www.jstor.org/stable/3133494>. Accessed 21 Sept. 2025.
- SONNINO, PAUL. "From D'Avaux to Dévot: Politics and Religion in the Thirty Years War." *History*, vol. 87, no. 286, 2002, pp. 192–203. JSTOR, <http://www.jstor.org/stable/24425636>. Accessed 21 Sept. 2025.
- Thaler, Peter. "Fall of the Peacemakers: Austria's Protestant Nobility and the Advent of the Thirty Years' War." *Renaissance and Reformation / Renaissance et Réforme*, vol. 39, no. 3, 2016, pp. 133–58. JSTOR, <http://www.jstor.org/stable/26398440>. Accessed 21 Sept. 2025.

Analyzing the Effects of Amman's Government Intervention On Fuel Consumption

By Omar Khoury

Essay question

To what extent has the Jordanian government intervention between 2019 to 2023 been successful in reducing the negative externality resulting from fuel consumption in the capital Amman?

Introduction

Amman is one of the most car-dependent cities in Jordan, with over half of its population relying on private transportation to commute. Jordan Times reported that the local fuel consumption in Amman is recently around 3,415 tons ("Local Consumption"). Decreasing fuel consumption is a common goal which is often publicly stated by the Jordanian government. This consequently incentivized them to impose an indirect tax on gasoline in January 2019, in order to decrease its consumption ("Government Makes Tax") but the indirect tax was ineffective due to gasoline having a low-price elasticity of demand (PED), which will be further elaborated in this essay. Therefore, In August 2022, the Jordanian government decided to subsidize public transportation methods such as the Amman bus as an alternative method of cutting fuel consumption within the city ("JD5m Allocated"). Since August 2022, the project has successfully transported 26 million passengers, and this number is expected to double after the second phase is completed ("PM Launches"). Jordan is ranked as one of the most polluted Arab countries, and ranked 46th globally, with a score of 50.7, according to the environmental performance index calculated by the Yale University ("Jordan Ranks"). This topic carries great significance, because the Jordanian government is currently considering options to address the rising pollution rates and its external consequences to society. The long-term consequences of pollution caused by fuel consumption, has to be assessed carefully to guarantee the adoption of the right policy in order to ensure **sustainability**. Therefore, this essay will demonstrate whether government intervention in Amman has been effective in reducing total fuel consumption by focusing on three key areas: first, the over consumption of gasoline in Jordan, second whether indirect taxes led to a decrease in fuel consumption; third, whether subsidies on electric vehicles contributed to a shift away from gasoline-powered cars.

Methodology

In order to effectively answer the research question, a variety of primary and secondary sources will be utilized such as surveys, interviews, observations, websites, academic journals, and government publications.

Primary data

- a) **Interview with the director of traffic operations in Amman:** This interview was another critical primary source, since it provided detailed information regarding

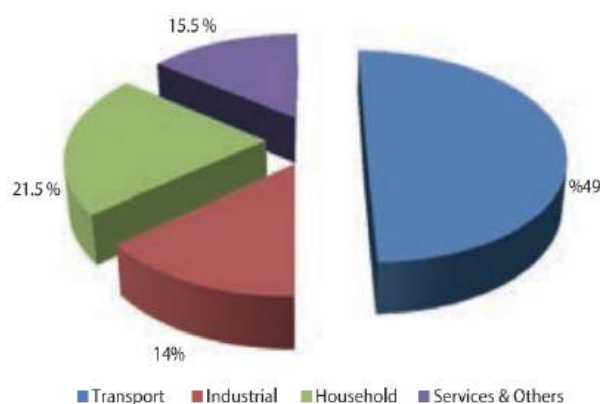
Kafaween's outlook on the Amman bus through in-depth questioning. In addition, the interview enabled me to obtain knowledge regarding the demand for the Amman bus throughout 2022 which was not present on the internet.

- b) Interview with the manager of the Gulf gas station:** This interview was yet another essential primary source, as it allowed me to obtain data that was not present on the internet, such as consumer responsiveness to the changes in gasoline prices. This helped me calculate the price elasticity of demand from 2019 to 2023. The findings led me to analyze the burden of the tax on various stakeholders.

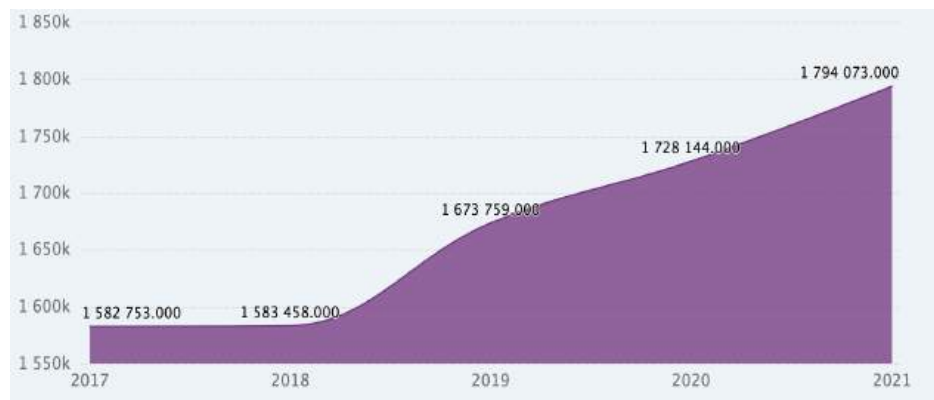
Secondary data

- a) Internet websites:** The internet is considered to be somewhat a reliable source of information for various aspects of study. It provides up-to-date information and thorough explanations. The data I collected from websites provided me with theoretical knowledge on market failure, externalities, and government intervention. However, unlike many countries, Jordan has limited amounts of fuel consumption statistics on the internet, therefore, I collected the needed information using national databases and government publications, such as documents released by the Ministry of Environment.

Body 1: Over consumption of gasoline in Jordan



The above chart illustrates the percentage of Amman's sectoral final energy consumption in 2019. The graph is divided into 4 main methods of energy consumption which are: transportation, industry, household, and services/others. From the pie chart it is clear that the majority of energy consumption is by transportation and household. Nearly half of total energy consumption in Amman is by transportation, accounting for 49%, while 21.5% is by households. Industrial consumption makes up 14% of total energy consumption. Only a small minority is caused by other sources and services; 15.5%. Therefore, proving that transportation is the main source of energy consumption in Amman.



This diagram demonstrates the significant increase in the number of cars in Amman between the years of 2017 to 2022. As shown from the diagram, the number of cars has increased from 167,3759 in 2019 to 172,8144 in 2020 and 179,4073 in 2021.

Percentage change in car consumption = $\frac{1794073 - 1673759}{1673759} \times 100 = 7.18\%$

This represents a 7.18% increase in car consumption over a span of 2 years.

In fact, cars represent 65% of the total vehicles in Amman and are responsible for all the national demand of gasoline fuel (Sawalha). Therefore, an increase in car consumption simultaneously increases gasoline consumption, being a complement, thus imposes threats on third parties, subsequently contributing to a market failure.

Market failure refers to inefficient allocation of resources in a free market. This therefore contributes to allocative inefficiency where too much or too little of the good is produced from the point of view of what is considered to be socially most desirable (Boyle). Gasoline is considered to be a demerit good, meaning that in Amman's market for gasoline, consumers are either unaware of its possible harms, or ignore the possible risks on themselves and on third parties thus causing a market failure (Blink and Dorton 130). This type of market failure, often referred to as a negative externality, occurs when the inefficient use of resources in the gasoline market results in a decrease in the welfare of society (As seen in figure 3 below).

The possible external costs of gasoline consumption include health issues such as lung cancer, asthma, bronchitis, work loss due to sick leaves, and the loss of productivity which may lead to unemployment. The environmental costs include air pollution, water pollution, and climate change consequently imposing threats on **sustainability**. Taking care of family members who were impacted by the emissions caused by gasoline require costly medical care, which may incur an opportunity cost where the resources and money spent could have been employed for another, more significant social benefit ("Gasoline Explained"). This created the need for government intervention, which incentivized the Jordanian government to impose a tax on fuel-powered cars.

Body 2: Weather indirect taxation led to a decrease in fuel consumption

The Jordanian government implemented an indirect tax of 0.165 JDS per one liter of gasoline in January 2019 (“Government Makes Tax”). This will decrease the supply of gasoline since the sales tax increases the cost for producing gasoline. Lower supply will therefore contribute to higher prices and In accordance with the law of demand, which states that "As price increases the quantity demanded will decrease ceteris paribus," car owners will be less able and willing to consume gasoline at higher price. If the tax level is set accurately and equal to the size of the negative externality, gasoline consumption is expected to decrease, reaching a socially optimal level. This strategy aims to internalize the externality resulting from gasoline consumption, and also make the external cost to society payable by the consumers. (Blink and Dorton 136). This method of **intervention** increased the Jordanian government’s tax revenue to 1.070 billion in 2019, and it is expected to rise by approximately 9.2% in 2023 (“Fuel Tax”). An increase in tax revenue will allow the Jordanian government to invest in medical care, education, infrastructure, and public transport thus contributing to economic growth in the long run. However, the outcome of the taxation just explained is the desired result which is far different from what frequently occurs. The impact of the indirect tax depends on many factors, but mainly gasoline price elasticity of demand (PED). Price elasticity of demand is a measure of how much the quantity demanded of a product changes when there is a change in the price of the product (Blink and Dorton 68). It is calculated using the following equation:

$$\text{PED} = \frac{\text{Percentage change in quantity demanded of the product}}{\text{Percentage change in price of the product}}$$

Following a change in price, the consumers' responsiveness differs depending on the price elasticity of demand of that product. If a product has inelastic demand, then the value of its PED is greater than zero and less than one. This signifies that a proportional change in the product's price results in a proportionally smaller change in its quantity demanded. Conversely, if a product has elastic demand, the PED value is greater than one and less than infinity, indicating that a proportional change in the product's price leads to a greater than proportionate change in its quantity demanded. Determining the PED of gasoline in Amman is highly significant because it shows whether the indirect tax will be effective. The number and closeness of substitutes is considered to be one of the main determinants of price elasticity of demand. Since gasoline has minimal close substitutes, then it will be difficult to reduce its total consumption following a change in its price. Consequently, suggesting that the demand for gasoline is relatively inelastic.

To effectively determine the price elasticity of demand (PED) for gasoline, I interviewed Omar Al Naber, the manager of the Gulf, a gas station in Amman, Jordan (Al Naber). During the interview, I asked him questions regarding the changes in price (JD) and quantity demanded of gasoline from January 2019 up until June 2023. The below table illustrates the data I have collected from the interview:

Month	PRICE /L JDS	SALES /L
Jan-19	0.695	7,450,554
Feb-19	0.695	7,076,977
Mar-19	0.72	8,124,704
Apr-19	0.75	8,192,727
May-19	0.75	8,157,884
Jun-19	0.75	8,263,768
Jul-19	0.75	9,064,582
Aug-19	0.775	8,677,883
Sep-19	0.755	8,075,155
Oct-19	0.77	8,202,984
Nov-19	0.765	7,723,424
Dec-19	0.775	7,781,039

The data presented in figure 5 indicates that, for the majority of the months in 2019, the price of gasoline remained relatively stable. Nonetheless, fluctuations in its quantity demanded were observed, primarily driven by external factors other than price. Notably, the most significant change in price occurred between March and April 2019. As a result, I will focus on these specific months to examine the price elasticity of demand for gasoline to further prove its inelastic demand using the mathematical approach bellow:

Using table 1, the PED of gasoline between March 2019 and April 2019 can be calculated:

Percentage change in Quantity demanded of gasoline: $\frac{8,192,727 - 8,124,704}{8,124,704} \times 100 = 0.84\%$

Percentage change in price of gasoline: $\frac{0.75 - 0.72}{0.72} \times 100 = 4.16\%$

PED: $\frac{0.84}{4.16} = 0.20$

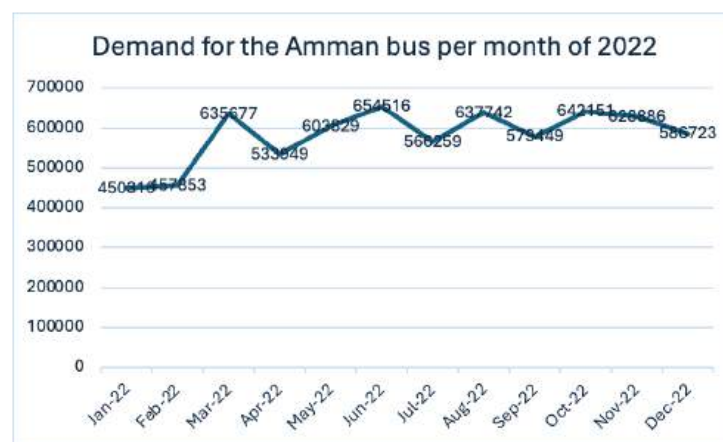
A PED of 0.20, suggests that demand for gasoline in Amman is relatively price inelastic. From this, we can conclude that a proportionate change in price will have a less than proportionate impact on its quantity demanded, meaning that the average consumer in Amman

will be relatively non- responsive to the changes in price of gasoline. Therefore, making indirect taxation an effective way in reducing gasoline consumption. Incentivizing the Jordanian government to look for other ways to reduce fuel consumption such as subsidies on public transport.

Body 3: subsidy on public transport

In August 2022, the government of Jordan implemented a JD 5 million subsidy on public transport, the amount of the subsidy ranged between JD 50 and JD 200 a month, depending on the route distance and the size of the vehicle (“JD5m Allocated”). By implementing this subsidy, the intention was to reduce production costs, and subsequently lower prices. This incentivizes producers to produce a larger quantity of public transport vehicles. Consequently, the Marginal Social Cost (MSC) curve would shift downward. This would result in a surplus of the Amman bus, prompting producers to decrease prices. With a lower price and greater quantity, quantity demanded would increase. This case illustrates how, in an ideal economic scenario, the introduction of subsidies can increase consumer demand, as outlined by economic theory. On the other hand, since the Jordanian government pays the subsidy, then it may be a burden on their budget which may lead to a budget deficit (which is when the government's expenditure exceeds their revenue from taxes). In fact, the Jordanian government's budget deficit in 2022 is equal to JD 1,291.6 million which is 5.9 percent of Jordan's GDP (*Recent Monetary*). This will prevent the Jordanian government from effectively dealing with emergencies and spending towards future generations. In addition, the subsidy may also incur an opportunity cost, where the money spent could have been used elsewhere for another significant social benefit, such as health care or education.

To effectively determine whether the 2022 subsidy on the Amman bus was effective, an interview was conducted with Ola Kafaween, the director of traffic operations in Amman (Kafaween). During the interview, questions were asked regarding the change in demand of the Amman bus from January to December 2022. The below graph summarizes the data collected from the interview:

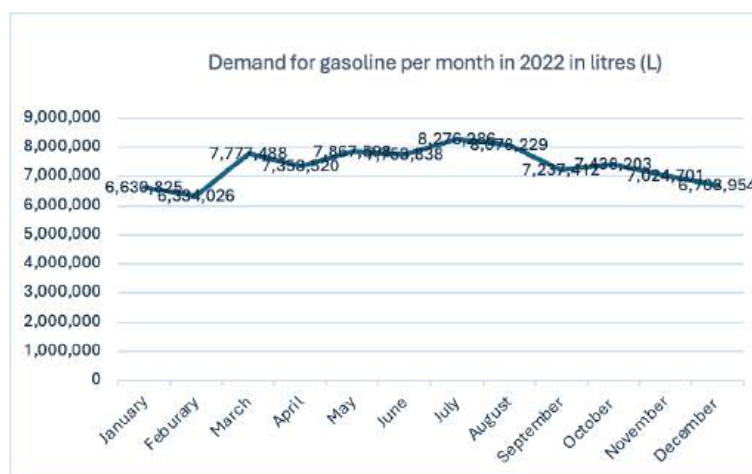


As revealed by the above data, the subsidy on the Amman bus in August 2022 caused a significant increase in its demand. As shown from figure 11, the demand for the Amman bus increased from 566,259 passengers in July 2022 (before the subsidy was implemented) to 637,742 passengers in August 2022 (after the subsidy was implemented).

$$\text{Percentage change in the demand of the Amman bus} = \frac{637742 - 566259}{566259} \times 100 = 12.62\%$$

This represents a 12.62% increase in demand for the Amman bus over a span of one month. Nonetheless, the data shows fluctuations in demand for the bus following the implementation of the subsidy, primarily driven by uncontrollable external factors. Despite the fluctuations, demand consistently remained higher than it was pre-subsidy in July 2022.

The findings prove that Jordan's government **intervention** has been successful in increasing consumer demand for the Amman bus by incentivizing consumers through a subsidy. Furthermore, given that the Amman bus uses Euro 5 diesel which releases 25% less emissions in comparison to car fuel, then an increase in its demand would simultaneously contribute to a decrease in fuel consumption in the city ("Limits to Improve")



As shown from the above graph, the subsidy on the Amman bus in August 2022 did not only cause a significant increase in demand for the bus, but also a decrease in demand for gasoline, as consumers would be incentivized to switch to the cheaper alternative. As shown from figure 13, the demand for gasoline decreased from 8,276,286 L in July 2022 (before the Amman subsidy was implemented) to 7,237,412 L in September 2022 (after the Amman subsidy was implemented).

$$\text{Percentage decrease in the demand for gasoline} = \frac{7237412 - 8276286}{8276286} \times 100 = -12.55\%$$

This therefore represents a 12.55% decrease in the demand for gasoline over a span of two months. Therefore, proving that the decrease in price of the Amman bus as a result of the subsidy, contributed to a decrease in demand for gasoline, consequently reducing fuel consumption in the city.

Conclusion

This essay has evaluated the effectiveness of Jordanian government intervention from 2019 to 2023 in reducing fuel consumption in Amman by examining the overconsumption of gasoline, the impact of indirect taxation, and the results of public transportation subsidies. Firstly, it was determined that transport is the largest energy consumer in Amman, and private vehicle use is directly linked to the consumption of gasoline. The rising motor vehicles and their reliance on gasoline created negative externalities, including air pollution and health risks, which was a case of market failure, this therefore called for government intervention. Secondly, the government's initial intervention through an indirect tax on gasoline in 2019 was aimed at reducing consumption. But since gasoline has an inelastic demand, as demonstrated by primary data from a gas station manager and computed PED of 0.20, the tax did not significantly decrease quantity demanded. This made indirect taxation ineffective in modifying consumer behavior. Third, a more effective solution was implemented in August 2022, in which the government imposed a JD 5 million subsidy on public transportation, specifically the Amman bus. Primary data from the director of traffic operations showed that the demand for the bus increased by 12.62% within a month. In addition, the demand for gasoline fell by 12.55% within two months, demonstrating the clear success of the subsidy in changing consumer behavior. The analysis demonstrates that while taxation was unsuccessful in reducing fuel consumption due to inelastic demand, subsidies on cleaner public alternatives achieved greater success toward both economic and environmental targets. This necessitates further research on optimal subsidy and incentive design in reducing negative externalities in car-dependent cities like Amman. Further research may examine the long-term behavior changes brought about by continued subsidies, or compare the success of similar interventions in other nations or cities with similar levels of car dependence and fuel usage. Techniques may include larger-scale surveys or econometric modeling in order to predict outcomes under a range of policy combinations.

Work cited

- Al Naber, Omar. Personal interview, 20 July 2025.
- Blink, Jocelyn and Ian Dorton. Economics Course Companion. Oxford University Press, 2020.
- Boyle, Michael. "Market Failure: What It Is in Economics, Common Types, and Causes". Investopedia, 25 July 2025, Accessed July 30 2025, <https://www.investopedia.com/terms/m/marketfailure.asp> .
- Final Energy Consumption by Sector 2019. Screenshot. Energy 2019 Facts and Figures, Jordanian Ministry of Energy and Mineral Resources, Accessed 29 July 2025, https://www.memr.gov.jo/ebv4.0/root_storage/en/eb_list_page/bruchure_2019.pdf .
- "Fuel Tax Revenues Increased from JD463 in 2015 to JD1.070b in 2019". Jordan News, 23 July 2025, , <https://www.jordannews.jo/Section-113/All/Fuel-tax-revenues-increased-from-JD463-in-2015-to-JD1-070b-in-2019-26327> .
- "Gasoline Explained. Gasoline and the Environment". Eia ,29 July 2025, Accessed 5 July 2023, <https://www.eia.gov/energyexplained/gasoline/gasoline-and-the-environment.php#:~:text=Gasoline%20use%20contributes%20to%20air%20pollution&text=The%20vapors%20given%20off%20when,carbon%20dioxide%2C%20a%20greenhouse%20gas>
- "Government Makes Tax On Fuel Fixed". Jordan News, 2 June 2019, Accessed 28 July 2025, https://petra.gov.jo/Include/InnerPage.jsp?ID=16187&lang=en&name=en_news#:~:text=The%20tax%20for%20both%20diesel,asphalt%20will%20be%20JOD45%2Fton.
- "JD5m Allocated as Subsidy for Public Transport Vehicles amid Rising Fuel Prices". Jordan Times, 3 August 2022, Accessed 5 August 2025, <https://www.jordantimes.com/news/local/jd5m-allocated-subsidy-public-transport-vehicle-s-amid-rising-fuel-prices>
- "Jordan Avg Monthly Wage Per Employee: Professionals". Ceic, Accessed 5 August 2025, <https://www.ceicdata.com/en/jordan/average-monthly-wage-per-employee/avg-monthly-wage-per-employee-professionals#:~:text=Jordan%20Avg%20Monthly%20Wage%20per%20Employee%3A%20Professionals%20data%20was%20reported,to%202020%2C%20with%2021%20observations> .
- Jordanian Families Spend 17% of Income on Transportation. Online Graph, "Jordanian Families Spend 17% of Income on Transportation". Jordan News, 5 August 2025, Accessed 3 August 2023, <https://www.jordannews.jo/Section-109/News/Jordanian-families-spend-17-of-income-on-transportation-says-World-Bank-17750#:~:text=It%20is%20divided%20into%20private,increased%20by%20only%2044.7%20percent.%E2%80%9D>
- Jordan Number of Registered Vehicles. Online graph , "Jordan Number of Registered Vehicles", Ceicdata Accessed 25 July 2025, <https://www.ceicdata.com/en/indicator/jordan/number-of-registered-vehicles>

The Future of Human Gene Editing Through CRISPR/Cas9 By Samita Golla

Keywords

CRISPR/Cas9, DNA, SgRNA, Ethics, Gene editing, Gene therapy

ABSTRACT

While originally discovered as a bacterial defense mechanism against invading bacteriophages, Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) has now become a promising new tool for gene editing. Using the Cas9 protein and a single-guide RNA for sequence specific targeting, CRISPR can be utilized to create double-stranded DNA breaks to edit targeted genes or DNA sequences in the genome. Using CRISPR/Cas9 gene editing can fix harmful genetic mutations and cure genetic disorders. Challenges associated with the use of CRISPR/Cas9 still remain such as difficulty delivering it to cells without triggering an immune reaction and off-target effects. Another issue that remains prominent is the ethics of using this technology; there are ongoing debates about in what cases it is ethical to use this gene editing tool on humans because of its potential for misuse. While the future of gene therapy using the CRISPR/Cas9 technology is hopeful, there remain many further discussions and necessary developments involving this tool.

INTRODUCTION

What is DNA?

[a] The Structure of DNA

DNA is a double stranded molecule that carries the entirety of a cell's genetic material. It is built up from nucleotides which are composed of a 5-carbon sugar (each carbon is numbered from 1' to 5' giving an orientation) called deoxyribose, a phosphate group, and one of four nitrogenous bases. These four possible nitrogenous bases are Adenine, Thymine, Cytosine, and Guanine. Structurally, DNA forms a double helix shape and the so-called "backbone" of this helix is made of alternating deoxyribose sugars and phosphate groups (**Figure 1**). The phosphate group is attached to the 5' carbon of the 5-carbon sugar and a hydroxyl group is attached to the 3' carbon of the sugar. Nucleotides link to each other through the 5' phosphate group binding to the 3' hydroxyl group of the previous nucleotide. The orientation of nucleotides is responsible for DNA directionality in the 5' to 3' direction, and since DNA strands are complementary to each other, the other strand runs antiparallel in the 3' to 5' direction. The other part of the DNA helix are the nitrogenous bases which are held together by hydrogen bonds. The base Adenine only pairs with Thymine and Cytosine only pairs with Guanine. This basepairing is only possible due to the antiparallel orientation of the strands. These base pairing rules are what makes DNA's two strands complementary to each other ¹.

[b] Cellular division requires DNA replication

The cell cycle consists of four stages: G1, S, G2, and M. A cell's life starts out in the G1 stage where it is mainly just growing and performing regular cellular functions. In the S stage, DNA replication occurs where the cell temporarily has double its original genetic material. The process of DNA replication is semi-conservative, in that the original doubled stranded DNA is split into two strands so that each new DNA helix has one newly synthesized strand of DNA and one old strand. In the DNA replication process, enzymes such as DNA polymerase use the parent strand as a template to synthesize a new daughter strand from each parent strand following the nitrogenous base pairing rules. DNA complementarity is directionality dependent, meaning that replication of the new DNA can only occur in the 5' to 3' direction ². The M stage is the mitotic stage and it involves the division of all the somatic cells (all body cells except the reproductive cells). The duplicated DNA made during the S stage is split up into two new identical cells in the process of mitosis. Any changes made to the DNA of a parent cell are passed down to the new daughter cell through mitosis, however changes made to somatic cells are not heritable to progeny for organisms that give rise to offspring through sexual reproduction. For example, a spontaneous mutation that occurs in a parent's somatic cell is not passed down to their offspring; for changes in DNA to be seen in offspring, mutations must be present in sex cells, egg and sperm in humans ³. This same concept of inheriting genetic changes applies to deliberate gene editing.

[c] DNA instructs cellular functions through transcription and translation.

DNA is the storage location for cellular instructions, and transcription is the first step to decode these instructions. Transcription is the process whereby messenger RNA (mRNA) is synthesized from a strand of DNA, which is done through the usage of several enzymes. The main enzyme, RNA polymerase, initiates transcription by binding to a promoter sequence on a specific gene. An example of a promoter sequence is a TATA box, a sequence that consists mainly of alternating thymine (T) and adenine (A) nucleotides. RNA polymerase then unzips the DNA double helix and begins attaching complementary bases of RNA to one strand of the DNA except it uses the base uracil instead of thymine when pairing with adenine to create mRNA.

Transcription stops when RNA polymerase runs into a terminator sequence, causing it to fall off of the template DNA strand and end transcription. The main difference between RNA and DNA is that RNA has the nitrogenous base uracil (U) instead of thymine and RNA has the sugar ribose instead of deoxyribose ².

Translation is the process of using mRNA as instructions to create a protein. The mRNA is read by the ribosome three base pairs at a time as a codon and each codon corresponds to one of the twenty different amino acids through the use of a tRNA molecule. Since there are 64 different possible codons, some codons correspond to the same amino acid. Each tRNA has a section on it called the anti-codon which is the complementary pairing of the codon on the mRNA (for example if a codon is AUG, then the anti-codon will be UAC). An important enzyme called aminoacyl-tRNA synthetases matches its specific amino acid to a tRNA molecule and as more amino acids are added, they form peptide bonds to each other. Translation stops when a

stop codon is encountered (UAG, UAA, or UGA). Then the bonded amino acid chain folds into the protein that the specific mRNA was coding for, often with the help of chaperon proteins ⁴.

[d] Mutations in DNA impact cellular function

Mutations in the DNA sequence can alter protein and cellular functions. The main types of mutations in DNA are deletions (nucleotides are deleted from a gene), insertions (nucleotides are inserted into a gene), inverses (a segment of DNA is cut out and reversed in orientation) or substitutions (where one or more nucleotides are swapped with another) ⁵(**Figure 2**). Insertions and deletions (indels), as well as inversions are referred to as frame shift mutations because these mutations change all the codon sequences that come after that mutation, changing most if not all of the amino acids in the protein, making them generally more impactful than substitution mutations. Substitution mutations alter one basepair within a codon which alters a single amino acid, and in the case of a “silent mutation” does not have any noticeable effect due to the fact that multiple codons correspond to the same amino acid. The real danger of substitution mutations though, is that they can sometimes cause a stop codon to appear earlier in the peptide sequence than it should, creating an unfinished protein. Any of these listed mutations can result in nonfunctional proteins which can cause detrimental disorders in people. Any changes to DNA sequences leads to changes to mRNA which then causes changes in proteins and their function in the cell⁶. However, with proper gene editing tools, deliberate changes in a DNA sequence can be introduced to elicit positive effects or restore original functionality.

RESULTS

What is CRISPR?

[a.] Background on CRISPR

CRISPR was first discovered in 1989 in *Escherichia coli* (E. Coli) bacteria as a defense mechanism against invading viruses called bacteriophages. CRISPR is made up of short repetitive sequences of DNA which are occasionally interrupted by unique spacers. These spacers are fragments of infecting viral DNA and they act as a “molecular record.” The recognition of a spacer fragment indicates to the bacteria there is a viral intruder and guides the destruction of that virus if it tries to invade the bacterium cell again. In addition to these sequences, an important component of the CRISPR complex is the Cas9 protein, which is responsible for cutting double stranded DNA at specific sites.

There are two different systems of CRISPR: Class I and Class II. Class I CRISPR utilizes multi protein complexes to find and cut DNA sequences, whereas Class II relies on a single multi-domain cas protein such as Cas9. Due to its efficiency and simplicity, the Class II system is primarily used for gene editing and engineering ⁷. There are different types of Cas9 proteins based on the strain of bacteria they come from. The main differences of these Cas9s are how immunogenic they are and specificity in their PAM sites, but in general they all have the same function of initiating the cut in the targeted DNA. SpCas9 (from the bacteria *Streptococcus*

Pyogenes) is often used for gene editing in CRISPR based therapies due to its efficiency in gene therapy. SaCas9 (from the bacteria *Staphylococcus Aureus*) is another important Cas9 protein for gene therapy; due to its small size, it is easier to deliver into human cells ⁸. Although SpCas9 was the first Cas9 to be employed for genomic editing, due to its large size and the fact that it causes immune responses in the body, researchers now use an engineered variant of SpCas9 for CRISPR projects that will be referred to simply as “Cas9” in this paper ⁹.

[b] The mechanism of CRISPR

The class II CRISPR system uses a guide RNA along with a Cas9 protein to identify and cut DNA for gene editing. A single guide RNA (sgRNA) is composed of two discrete sequences within the RNA molecule that are joined by a linker loop; the two sections are the trans-activating CRISPR RNA (tracrRNA) and CRISPR RNA (crRNA) ¹⁰. The crRNA contains spacers which are specific sequences of genes that the Cas9 protein can target, which can come from foreign DNA from viruses such as in the case of bacteria. The tracrRNA binds to the repeat regions of the crRNA which forms the sgRNA that the Cas9 protein then binds to ¹¹. For the sgRNA to be able to bind to its target gene, the Cas9 needs to find a Protospacer Adjacent Motif (PAM) site. A PAM site is a sequence of DNA starting with N(any nucleotide)GG (**Figure 3**). This site has to be present 3-4 bases downstream of the desired gene for CRISPR to work ¹¹. The tracrRNA starts the actual gene editing process by helping activate the Cas9 protein in cleaving the target DNA. The Cas9 protein has two endonuclease domains: HNH and RuvC. HNH cleaves the DNA strand that is complementary while the RuvC cleaves the non-complementary strand of DNA, causing a double stranded DNA break ¹². When it comes to using CRISPR in humans as an actual gene therapy, a specific sgRNA can be synthesized by researchers for precise targeting. The spacer sequences in the crRNA can be synthetically designed using a specific gene sequence so that the Cas9 protein will target a desired gene or sequence of DNA ¹¹.

There are two different types of repair mechanisms for double stranded DNA breaks: non-homologous end joining (NHEJ) or homology directed repair (HDR). NHEJ is a much less accurate repair system, as this process quickly re-connects broken DNA ends without a guide template. NHEJ prioritizes speed over accuracy so it often causes many mutations such as insertions or deletions in the DNA it tries to repair. The introduction of mutations, as previously discussed, often renders the protein that the gene was meant to code useless. HDR however, uses a template to properly repair the DNA break. If a modified DNA template is provided for HDR, it will direct the cell to remake the DNA sequence in the desired way. For example, with the right template HDR can fix a mutated gene by inserting the right nucleotides. HDR is a key way to use CRISPR for gene therapy because it can add a gene sequence to a DNA break site to restore or alter the function of a gene ¹³.

[c]. What CRISPR can do that other gene editing tools can not

Gene editing is a process that modifies one's genome to introduce beneficial changes to a person's DNA to change the original coding. Before CRISPR, a few prominent gene therapies

included Zinc-Finger Nucleases (ZFN) and Transcription Activator-Like Effector Nucleases (TALEN). ZFN and TALEN both use a specifically engineered protein that directly recognizes a DNA sequence, so each protein construct can only work at one specific site until they are re-designed. Whereas for CRISPR, multiple sgRNAs can be utilized to have Cas9 make multiple cuts. This makes CRISPR much more efficient than ZFN and TALEN gene editing which requires precise and time consuming protein design to achieve targeting that rapidly designed sgRNA in CRISPR can provide ¹⁴.

Furthermore, the CRISPR/Cas9 system can be much more modular and adaptable than TALENs and ZFNs as it can be used for a variety of other mechanisms to edit or alter genes. By modifying a Cas9 protein by fusing it with a deaminase enzyme, CRISPR/Cas9 can alter base pairs but without causing a double-stranded break. When the modified Cas9 and sgRNA bind to the target DNA a single-stranded DNA region called an R-loop forms. Then the deaminase enzyme attached to the Cas9 can chemically alter a nucleotide to another (for example changing a Cytosine into a Guanine). This form of CRISPR is called prime editing. As substitution mutations account for over 50% of genetic mutations ¹⁵, being able to fix these mutations could cure many genetic disorders and save lives, a feature that other genetic tools do not have. Combined with its adaptability and ability to not have to be re-engineered for every new DNA sequence CRISPR possesses advantages over other genetic tools that support the time spent researching and perfecting it for human trials.

[d]. Approaches to delivering CRISPR to human cells

There are different ways of delivering CRISPR/Cas9 components to human cells for gene therapy. The first is the ex vivo approach where cells are edited with CRISPR/Cas9 outside of the body in a culture medium and then those cells are re-planted back into the patient ¹⁶. An example of the ex vivo approach is the study by Everette et al. which edited Hematopoietic Stem and Progenitor Cells (HSPCs-cells that give rise to blood cells) from patients with Sickle Cell Disease¹⁷. Sickle Cell Disease is caused by a single point mutation in the HBB gene that produces abnormally shaped red blood cells that can not properly carry oxygen in the body. Everette et al. performed ex vivo prime editing on the extracted HSPCs to correct the point mutation and then the edited cells were transplanted into immunodeficient mice. After 17 weeks, most of the red blood cells made by the corrected HSPC cells in the mice functioned normally ¹⁷. The ex vivo approach has appeal in that it causes less of an immunogenic response than in vivo because the CRISPR editing tools that would cause that response are not seen by the host immune system ¹⁷. There are problems with the ex vivo approach however, it only works for cells that are easy to collect, can survive in the lab, and be able to function after being returned to the patient. There are a few cells that fit these criteria such as blood and immune cells but many others that do not such as neurons, making ex vivo CRISPR mostly limited to treating immune and blood disorders ¹³. Even with these shortcomings, the ex vivo approach is still valuable for researchers to pursue because there are still many blood and immune disorders that could benefit from being treated with CRISPR.

The second approach is in vivo delivery where the Cas9 and sgRNA are introduced directly into the body to enter cells; however, there are some issues with this approach such as immunogenicity that will be discussed later. One way of delivering CRISPR components in vivo to cells is through the use of plasmid DNA. The plasmid DNA carries the sequence of both the sgRNA and Cas9 protein and is introduced into the cell through viral delivery systems such as Adeno-Associated Virus (AAV) or non-viral systems like lipid nanoparticles (LNP), which will be discussed as well. Once in the nucleus, the plasmid DNA is transcribed into the sgRNA and an mRNA which is then translated into the Cas9 protein. In this way CRISPR components are delivered to the site of activity, inside a cell, without immune recognition. Some problems with this method include the requirement for a target cell to be actively transcribing and translating, not to mention the time for the cell to carry out those functions ¹⁶.

Another way of delivering CRISPR components to cells is by directly inserting (with a non-viral or viral delivery system) Cas9 mRNA and sgRNA so that the mRNA can be translated into Cas9 which will then bind with the sgRNA and begin gene editing. This method is shorter than using plasmids because it does not rely on both cellular transcription and translation to produce the CRISPR/Cas9 components. However, this method may be less stable because the delivered mRNA is often broken down by RNase enzymes in the cell. The last method is ribonucleoprotein (RNP) delivery where the sgRNA and Cas9 protein complex are delivered directly into the cell. Editing can start right away so this method is quickest, but due to the RNP's large size, it is hard for it to cross the cell membrane ¹⁸. However, all three of these methods using the in-vivo approach share some common problems.

[e]. Immunogenicity Problems In Using CRISPR As A Therapy

Early gene therapy treatments in humans led to serious side effects including patient death, leading to a slowdown in the development and application of gene therapies. Even though immunogenicity against CRISPR/Cas9 is still a huge problem when it comes to using it on patients, using CRISPR/Cas9 is becoming prominent in science fields.

Because CRISPR originates from bacteria, the human immune system can perceive the Cas9 protein as a foreign substance, triggering an immune response. In immune responses, immune cells in the body called B cells can produce antibodies specific for the foreign CRISPR components which bind and mark them for destruction. Due to CRISPR's bacterial origins, many people have already encountered the same bacteria CRISPR comes from and will have pre-existing antibodies for it, causing cells to recognize and destroy the therapeutic CRISPR components like it would a pathogen, before the gene editing process can even begin¹⁵.

MHC (Major Histocompatibility Complex) Class I molecules are proteins found on the surface of almost every cell in an individual's body. Their role is to show killer CD8 T cells (a type of immune cell that protects the body from pathogens) tiny protein fragments sampled from within the cell; in the case of a cell infected by a virus, MHC-I could present viral proteins to signal it is infected. If the killer T cell recognizes those fragments to be foreign, they may attack the presenting cell and induce an immune response ⁸. Raghavan et al. determined that a specific

region of Cas9 was displayed as an epitope on MHC-I to the killer T cells; the researchers mutated the Cas9 protein so that the MHC-I was no longer able to bind and present those fragments to prevent the killer T cells from being alerted of a foreign presence. The experiment aimed to make Cas9 less detectable by the immune system and still keep its DNA cutting ability intact ⁸. Overall the experiment showed a reduced immune response with the engineered Cas9; however, this process has only been tested in the lab and has not yet been used in patient trials so there is still a long way to go with addressing immunogenicity problems with CRISPR. While scientists are not there yet, in the future CRISPR could utilize a fully synthetic version of Cas9, not obtained from bacteria, to stop Cas9 specific immune responses.

The sgRNA can trigger immune responses as well due to certain chemical features like a 5' tri-phosphate group that are not normally found in human cell processes. Guide RNAs can be chemically altered or engineered to remove these phosphate groups, making them less likely to trigger an immune response¹⁵. Although sgRNA modification seems to be promising in reducing immune responses from the body, this method does not guarantee that the sgRNAs will be fully hidden from the immune system just yet and sgRNA immunogenicity remains a challenge. All in all, it is important to consider that it's not just Cas9 and RNAs that cause immunogenicity problems, it is also the vehicles used to deliver them to cells in the body. Viral vectors such as AAV were considered ideal for delivering CRISPR/Cas9 to cells because viral capsids (and the derived viral vectors) are naturally suited to release their contents into cells, thus they could be utilized to properly deliver CRISPR/Cas9 into cells for gene editing. However, since AAV is a virus, the immune system can recognize the viral proteins on the AAV and trigger an immunogenetic response, which in some cases has led to clinical trial patients experiencing serious adverse effects including death when an AAV delivery system was used. An alternative to AAVs are non-viral delivery systems such as LNPs which are spherical based lipid carriers that are designed to safely carry CRISPR components to cells, by protecting them from being degraded by enzymes in the body. LNPs are absorbed into the cell when they come in contact with the cell membrane and during that process the LNPs are destabilized, releasing the CRISPR components safely into the cell. LNPs are a promising delivery mechanism for CRISPR/Cas9 due to low immunogenicity ¹⁹.

[f.] Off Target Effects are a cause for concern with CRISPR

An important consideration when using CRISPR as a gene therapy is to ensure there are no off target effects. When it comes to binding to the target DNA sequence, it is expected that the sgRNA has been designed to bind to one exact complementary sequence, however, the sgRNA can tolerate a few mismatches and still “recognize” a target sequence. Thus, it is possible that the sgRNA could bind to a very similar sequence elsewhere in the genome. If this occurs, Cas9 will be directed to cleave the DNA at an incorrect site and cause a double stranded DNA break, potentially leading to indels in the wrong part of the genome, something like this can have a significant impact on the organism. Alterations to just one gene can lead to dangerous changes in cellular function and even cell death. Another undesired effect of CRISPR can actually happen at

the correctly targeted site. As mentioned prior, double-stranded breaks can be repaired by the cell incorrectly, in a harmful way that causes large genomic mutations such as deletions, insertions, duplications, or translocations (when a segment of DNA breaks off and attaches in the wrong area). Together, all these different mutations are called structural variants. Structural variants can sometimes cause oncogenes (mutated genes that can promote cancer growth) depending on the size of the structural variant and the type of cell it occurred in ²⁰.

One method being used to reduce the impact of off target effects is called CRISPR GUARD. GUARD is shortened sgRNAs that are complementary to the off-target sites in the genomes, so they occupy those off-target sites, leaving only the correct site open for the sgRNA to bind. Since the GUARD RNAs are a couple of base pairs shorter than a sgRNA, the Cas9 can bind to it but it cannot activate its cleaving abilities on the GUARD RNAs, protecting the cell from off-target effects. CRISPR GUARD has successfully protected genes from off-target binding, but it has not been tested in patients yet ²¹. In the future, GUARD RNAs can be incorporated into all CRISPR gene editing procedures to eliminate any potential off-target genomic alterations and overall make CRISPR safer to use on people.

Discussion

CRISPR is a promising tool for human gene editing because it has shown some success in pre-clinical and clinical trials to edit the genome in both ex vivo and in vivo applications. There are three main applications human germline editing through the use of CRISPR is expected to be used for. The first is preventing serious gene disorders that would be passed from parent to child (such as sickle cell anemia), the second is reducing the risk of common diseases in people such as heart disease, and lastly enhancing human abilities beyond what humans are capable of now.

While there are still issues of immunogenicity and off-target effects, solutions are being developed in the lab that have begun to show beneficial results. However there remains the unresolved and outstanding issue regarding the ethics of using CRISPR in humans which is not something that can be solved with experimentation alone. Gene editing has been around for decades and the ethical debate surrounding its use has been ongoing, but with the emergence of CRISPR/Cas9 and the ease of which gene editing can occur, there has been a re-emergence in ethical debate. Now, there is a renewed focus on whether or not human germline editing is ethical because those edits could become heritable in future generations. In 2015, the first international summit on human gene editing took place, which came to the conclusion that before anyone should consider human germline editing, scientists must prove that it can be done safely and accurately. Furthermore, the summit concluded that the use of this technology should have general public agreement not just based on the decision of a few people. In 2018, a second conference on human gene editing was meant to take place but before it could, news broke that He Jiankui used CRISPR to create genetically altered babies which broke the conditions agreed upon in the 2015 conference ²².

In 2018, Jiankui He performed germline genomic editing using CRISPR to modify the

CCR5 gene in two twin babies with the hypothesis that this edit would render the babies immune to HIV by removing the CCR5 protein receptor that the HIV virus uses to bind to and enter cells. This study involved an HIV positive father and an HIV negative mother -these conditions made the risk of actually transmitting HIV to a baby low. His study was met with global criticism due to his insufficient medical experience and education. Furthermore, the genomes of the babies edited were not likely to be born with HIV anyways, meaning this experiment could have caused harmful mutations to the genomes of HIV negative babies.

A few months after He's study, Russian scientist Denis Rebrikov planned a similar experiment and in October of 2019 was reported to be working on editing the GJB2 gene which is known to be associated with deafness. The ministry of Health of the Russian Federation denied Rebrikov's appeal to perform this study, citing that the work could end up being disastrous, effectively putting an end to Rebrikov's planned study ²³. Due to Rebrikov planning his study recently after the controversy He caused with his study, Rebrikov's plan received backlash and scrutiny causing his study to be ended. The halting of this study could be considered a win for the ethical use of CRISPR because another potentially unethical human trial study of CRISPR was stopped.

Despite the recent progress in CRISPR/Cas9 development, the views on human germline editing remain mixed. Use of CRISPR/Cas9 for germline editing can be viewed positively when the intent is to prevent individuals from having lifelong genetic diseases, such as sickle cell anemia and cystic fibrosis, thereby greatly improving quality of life or even saving lives where such disorders are lethal. On the other hand, human germline editing may not qualify as a protective genetic therapy if the edits are not preventing diseases but rather altering subjectively unfavorable traits. A condition like asthma could be argued both ways, while it is not life-threatening curing asthma can definitely improve a person's quality of life.

The biggest debate in human germline editing lies in using it for genetic enhancement. This means editing the genome for traits that do not directly correspond to an individual's health such as blue eyes or brown hair ²². Over time, doing this can erase diversity and cause a future where babies are genetically designed to have particular phenotypes. This raises the question of if human germline editing is a way to save lives or if it is just a way to create "designer" babies. While CRISPR/Cas9 is a good technology and gene editing alone is not necessarily a bad tool, it can be used for morally grey purposes. Therefore, a line needs to be drawn between using CRISPR/Cas9 to prevent/cure diseases and improving quality of life to using it for subjective trait enhancement.

Although CRISPR has not been perfected and remains under debate as to what ethical boundaries should be set in place for gene editing, it remains an important tool that is currently being used in clinical trials with the potential to one day change the medical field for the better and save countless lives.

Acknowledgements

Special thanks to my mentor Brett Higgins for all the help in making this paper

Funding

No Funding to declare

Declaration of Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication of this article.

Works Cited

1. Alberts, B. et al. The Structure and Function of DNA. in *Molecular Biology of the Cell*. 4th edition (Garland Science, 2002). ISBN 10: 0815332181
2. Mercadante, A. A., Dimri, M. & Mohiuddin, S. S. Biochemistry, Replication and Transcription. in StatPearls [Internet] (StatPearls Publishing, 2023).
3. Alberts, B. et al. Mitosis. in *Molecular Biology of the Cell*. 4th edition (Garland Science, 2002). ISBN 10: 0815332181
4. Alberts, B. et al. From RNA to Protein. in *Molecular Biology of the Cell*. 4th edition (Garland Science, 2002). ISBN 10: 0815332181
5. Mahmoud, M. et al. Structural variant calling: the long and the short of it. *Genome Biology* 20, 246 (2019).
6. Ghannam, J. Y., Wang, J. & Jan, A. Biochemistry, DNA Structure. in StatPearls [Internet] (StatPearls Publishing, 2023).
7. Asmamaw, M. & Zawdie, B. Mechanism and Applications of CRISPR/Cas-9-Mediated Genome Editing. *Biologics : Targets & Therapy* 15, 353 (2021).
8. Raghavan, R. et al. Rational engineering of minimally immunogenic nucleases for gene therapy. *Nature Communications* 16, 1–13 (2025).
9. Kovalev, M. A., Davletshin, A. I. & Karpov, D. S. Engineering Cas9: next generation of genomic editors. *Applied Microbiology and Biotechnology* 108, 209 (2024).
10. Liu, Y., Liu, W. & Wang, B. Engineering CRISPR guide RNAs for programmable RNA sensors. *Biochemical Society Transactions* 51, 2061 (2023).
11. Cong, L. et al. Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science (New York, N.Y.)* 339, 819 (2013).
12. Jinek, M. et al. A programmable dual RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (New York, N.Y.)* 337, 816 (2012).
13. Palacios, A. M., Korus, P., Wilkens, B. G. C., Heshmatpour, N. & Patnaik, S. R. Revolutionizing in vivo therapy with CRISPR/Cas genome editing: breakthroughs, opportunities and challenges. *Frontiers in Genome Editing* 6, 1342193 (2024).
14. Li, T. et al. CRISPR/Cas9 therapeutics: progress and prospects. *Signal Transduction and Targeted Therapy* 8, 1–23 (2023).
15. Kim, H. S., Kweon, J. & Kim, Y. Recent advances in CRISPR-based functional genomics for the study of disease-associated genetic variants. *Experimental & Molecular Medicine* 56, 861–869 (2024).
16. Ewaisha, R. & Anderson, K. S. Immunogenicity of CRISPR therapeutics—Critical considerations for clinical translation. *Frontiers in Bioengineering and Biotechnology* 11, 1138596 (2023).
17. Everette, K. A. et al. Ex vivo prime editing of patient haematopoietic stem cells rescues sickle-cell disease phenotypes after engraftment in mice. *Nature Biomedical Engineering* 7, 616–628 (2023).
18. Lee, H., Rho, W.-Y., Kim, Y.-H., Chang, H. & Jun, B.-H. CRISPR-Cas9 Gene Therapy:

Non-Viral Delivery and Stimuli-Responsive Nanoformulations. *Molecules* **30**, 542 (2025).

19. Taghdiri, M. & Mussolino, C. Viral and Non-Viral Systems to Deliver Gene Therapeutics to Clinical Targets. *International Journal of Molecular Sciences* **25**, 7333 (2024).
20. Hunt, J. M. T., Samson, C. A., du Rand, A. & Sheppard, H. M. Unintended CRISPR-Cas9 editing outcomes: a review of the detection and prevalence of structural variants generated by gene-editing in human cells. *Human Genetics* **142**, 705 (2023).
21. Coelho, M. A. et al. CRISPR GUARD protects off-target sites from Cas9 nuclease activity using short guide RNAs. *Nature Communications* **11**, 1–12 (2020).
22. Almeida, M. & Ranisch, R. Beyond safety: mapping the ethical debate on heritable genome editing interventions. *Humanities and Social Sciences Communications* **9**, 1–14 (2022).
23. Gostimskaya, I. CRISPR–Cas9: A History of Its Discovery and Ethical Considerations of Its Use in Genome Editing. *Biochemistry. Biokhimiia* **87**, 777 (2022).

Figures

Figure 1: The structure of DNA. The figure shows how in the basic structure of DNA, each phosphate group (blue circle) is attached to the 5-carbon sugar deoxyribose (orange pentagon) which in turn is attached to a nitrogenous base. These 3 components make up a nucleotide. Nitrogenous bases connect to their corresponding nitrogenous base by hydrogen bonds, (represented as dotted lines) connecting several nucleotides in the double helix shape of DNA.

Figure 2: DNA mutations change mRNA which can change proteins. (a) Normal: A sequence of DNA that shows the corresponding mRNA (in red) and the specific amino acid the codon calls for (b) Substitution Mutation: A sequence of DNA that shows a Cytosine nucleotide instead of a Guanine nucleotide making the corresponding codon to call for a stop, ending the protein early (c) Silent Substitution Mutation: A sequence of DNA that shows a Adenine nucleotide instead of a Guanine nucleotide, but since both of these codons call for the amino acid Tyrosine, the mutation is silent and the coded protein stays the same (d) Insertion Mutation: A sequence of DNA with an extra Adenine nucleotide inserted between the two Guanine nucleotides. This insertion leads to a Serine amino acid being called for instead of an Arginine and leaving an extra Adenine nucleotide, shifting all the codons that come after (e) Inversion Mutation: A sequence of DNA that shows the ATG nucleotide sequence being inverted and shown as GTA which codes for the Leunine amino acid instead of the correct Tyrosine amino acid (f) Deletion Mutation: A DNA sequence that shows the Guanine nucleotide between the Thymine and Guanine nucleotide being removed from the sequence. Even though this still causes the corresponding codon to call for the Tyrosine amino acid, the Cytosine and Adenine nucleotide are left over, shifting the codons that come after

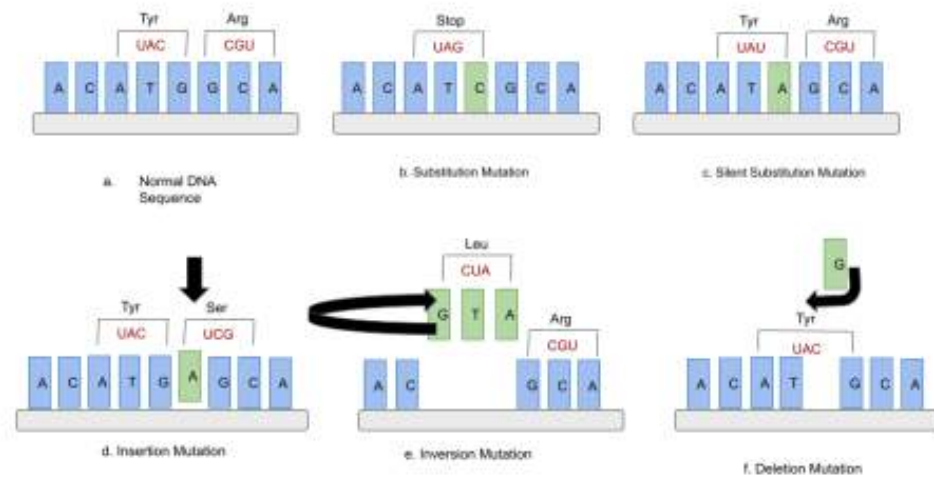
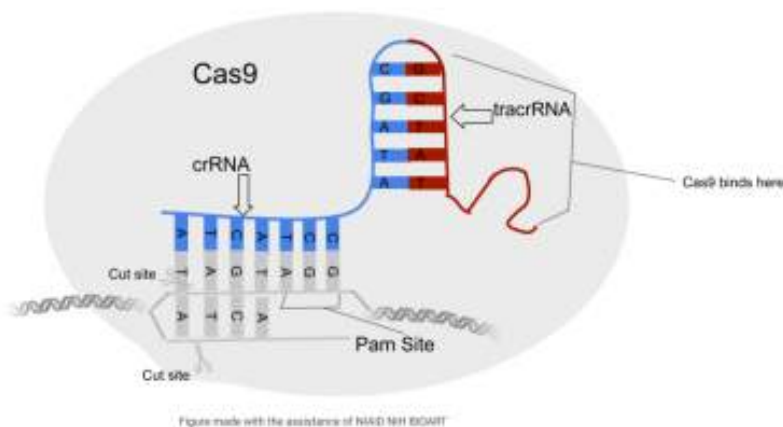


Figure 3: SgRNA + Cas9 form a complex for DNA editing. The figure shows a single guide RNA (sgRNA) which is composed of a crRNA (CRISPR RNA), which binds the targeted DNA sequence and a tracrRNA (trans-activating CRISPR RNA), which binds and activates the Cas9 protein. The sgRNA guides a bound Cas9 protein to a complementary sequence of DNA which must have a series of nucleotides called a PAM (Protospacer Adjacent Motif) site a few basepairs down from the targeted DNA sequence for effective binding. Once bound, the Cas9 protein initiates double-stranded DNA cleavage.



From Tool to Crutch: The Thin Line in AI-Enhanced English Learning By Siqi Fan

Introduction

In today's high school English classrooms, artificial intelligence (AI) has become almost as common as textbooks. It streamlines the learning process and alleviates the burden on teachers, enabling them to transition from routine tasks such as error correction to more substantive interactions that promote critical thinking and provide personalized instruction. Furthermore, AI acts as a flexible learning aid, integrating smoothly into students' everyday English practice and delivering tailored support across multiple language competencies. A systematic review by Daud et al. (2025), which encompasses 89 studies on high school students, offers a comprehensive overview of AI integration in English language teaching. It corroborates the significant advantages of AI in terms of personalization and efficiency, but also consolidates extensive evidence of its potential drawbacks, including over-reliance and the weakening of critical thinking skills. This analysis establishes a foundation for examining the dual role of AI in educational contexts. According to Daud et al. (2025), grammar checkers can rapidly identify neglected errors and spelling mistakes, consistent with their conclusion that writing assistance is the most frequent application of AI. However, the authors also emphasize that although AI enhances the efficiency of writing practice, it may lead to excessive dependency, impairing students' capacity for independent and critical thought. This paradox of AI as both a facilitator and a potential hindrance lies at the core of contemporary English language education. Teachers, as guides in this evolving landscape, play an essential role in navigating this tension.

This paper will address two central questions: (1) How do AI tools influence students' learning processes and outcomes in English classes? (2) How can teachers and students maximize the benefits of AI while mitigating its negative effects? The findings are intended to provide practical recommendations for balancing AI tools with effective English instruction, assisting educators and learners in integrating technology more thoughtfully into the classroom.

1.How students use AI for English Learning

Students are increasingly turning to AI tools for English learning, a trend that has led to growing concerns about over-reliance (Daud et al., 2025).

1.1 Using AI tools for speaking practice

AI has shown notable effectiveness in enhancing students' oral English skills. In an empirical study by Yang et al. (2025), 308 Chinese EFL undergraduates participated in an experiment examining how different interaction styles of AI peer agents influenced their speaking practice. The study evaluated several dimensions: oral proficiency (including pronunciation, fluency, and coherence), cognitive load, learning anxiety, and technology acceptance. Data were collected through performance assessments, surveys, and questionnaires to comprehensively capture the impact of AI interactions on speaking development. To assess

these areas, they evaluated oral skills like pronunciation, fluency and coherence of ideas. They also used questionnaires and surveys to measure cognitive load, learning anxiety and technology acceptance. In their study, EFL university students practiced with AI “virtual peers”. When using “virtual peers” chatbots for speaking practice, even the shyest ones began to speak more with confidence. Surprisingly, the study found that this group of students improved their oral English skills by 23%, with more significant progress in pronunciation and coherence of ideas. It is evident that reduced anxiety is followed by accelerated learning speed. Similarly, Hu et al. (2025) conducted a meta-analysis to evaluate generative AI’s impacts on student learning outcomes, systematically synthesizing, analyzing, and combining data from relevant existing studies. Their analysis found that AI helps some students a lot, such as those who use AI tools to practice speaking for hours continually, and this is because AI creates a relaxed environment without judgment, letting students speak freely without consciousness. It provides instant feedback on pronunciation or grammar, allowing them to make quick adjustments. It also adapts content to students’ own level to keep practice engaging, which helps them stay focused for hours.

Nonetheless, it confuses students who get overwhelmed by too many options from AI tools. For example, AI-based language learning platforms always offer a wide range of practice modules, like vocabulary quizzes with multiple difficulty levels, grammar exercises categorized by tenses, and speaking tasks with various topics. When faced with such choices, students may struggle to decide which one to focus on, they may waste time in endless browsing instead of engaging in effective practice. Some may even feel anxious about missing out on better options, which impairs their learning efficiency. Additionally, Yang et al. used AI “peers” for speaking practice. These are chat interfaces with pre-programmed conversation starters. They follow fixed language rules for responses and lack real-time emotional cues such as changes in tone or nonverbal signals. Their research found that these AI tools alleviate students’ anxiety about being mocked in class. However, interactions with AI feel rigid and lack the feeling of face to face conversations with real people. Daud et al. also pointed out that while AI has its merits, it cannot recreate the interpersonal elements of learning, such as working together to correct mistakes or sharing a laugh over mispronunciations.

1.2 Using AI tools for writing tasks

High school students are increasingly using AI tools as writing assistants, a trend observed by Lai (2025). He watched how students interacted with AI tools, compared writing done with AI help to work done on their own, looked for weaknesses in critical thinking and original expression, and used surveys and structured interviews to figure out why students rely too much on AI. His research shows that this kind of reliance often makes high school students less willing to try hard on their own. They tend to depend on suggestions from AI instead of creating sentences or analyzing texts by themselves. Plus, the research found that many seek to avoid the discomfort of getting feedback from peers or teachers, which is quite similar to how some avoid feedback in speaking practice. Their growing dependence on AI is mainly because

these tools are fast and easy to use.

Additionally, AI tools such as ChatGPT have also shown its value in EFL writing instruction for high school students by offering benefits for skill development. Pei et al. (2025) tested this with 99 Grade 10 students in a Hong Kong secondary school over 10 weeks, dividing them into two groups—The experimental group used ChatGPT for core tasks, including generating outlines, revising grammar, and expanding ideas. By contrast, the control group relied on traditional methods. These methods included grammar workbooks and "peer review", a process where students exchange drafts to provide feedback. The ChatGPT group scored 15% higher in overall writing quality, with 21% fewer grammar errors and 18% better paragraph flow. Notably, it benefited students across proficiency levels that low-achieving students (initial scores below 60/100) improved by 28%, and high-achieving ones (above 80/100) by 9%, suggesting AI closes skill gaps through targeted support. In addition, AI can enhance writing assessment. Topics et al. explored AI's role in grading by comparing automated essay scoring (AES) tools with 30 teachers on 500 student essays. The assessment they used focused on three key aspects—grammar, organization, and vocabulary. The audience here is the students who wrote these 500 essays, though their specific grade level is not mentioned. Results showed AES aligned with human ratings 82% of the time for grammar, 79% for organization, and 76% for vocabulary. This consistency totally reduced teachers' grading time by 40%, allowing more individual meetings to discuss content which further boosts writing skills.

1.3 AI tools help students work more efficiently

AI tools also enhance efficiency for students facing academic pressure, yet this often exacerbates excessive reliance. Hu et al. conducted a meta-analysis, combining results from several studies to assess how generative AI affects student learning outcomes. They showed that heavy workloads and tight deadlines leave students with little time for intentional language practice. This leads students to use AI to finish tasks more quickly, whether for translations, comprehension exercises, or script drafting. Nevertheless, this kind of convenience harms students' meaningful learning. When they rely on AI to make tasks easier, they skip critical thinking for gaining knowledge, such as processing information and refining language. Hu et al. further stated that focusing on AI tools for efficiency breeds a habitual reliance. It lowers students' participation in learning and weakens their language proficiency over time.

1.4 Using AI tools for reading comprehension

AI tools, particularly those offering summarization support have demonstrated considerable potential in enhancing reading comprehension. In a study conducted by Huang et al. (2025), 163 university sophomores from various disciplines in Taiwan were examined to assess the learning effects of AI summary tools. The findings indicate that such tools significantly improve comprehension and information retention when students actively engage with the output. For instance, learners who reformulated AI-generated summaries rather than copying them passively showed better recall of key textual details. This process of active reconstruction

encourages deeper interaction with the material, thereby strengthening conceptual understanding and long-term retention. However, the study also identified risks associated with passive consumption of AI summaries. Merely copying AI-generated content was found to inhibit learning, as it reduces cognitive engagement, such as identifying main ideas, making connections, and synthesizing information which are essential for developing robust reading skills.

1.5 Social acceptance of AI in education

Growing social acceptance of AI in educational contexts is further reinforcing student dependence on these tools. Lampropoulos et al. (2025) analyzed over 12,000 ChatGPT-related discussions on Twitter and found that both educators and students increasingly perceive AI as an essential and legitimate learning aid. This normalization, however, often obscures its limitations. The same study revealed that ChatGPT struggles with nuanced language elements such as sarcasm and cultural references—a shortcoming frequently noted by educators. More concerning is the observed phenomenon of unconditional trust that users often accept AI-generated content without critical verification. These factors—broad social acceptance, unrecognized system limits, and uncritical usage collectively deepen students' over-reliance on AI. While student adoption patterns shed light on evolving learning behaviors, teachers' motivations and instructional approaches also play a critical role in shaping how AI is integrated into the classroom.

2. Why teachers use AI in English classes

In recent years, the integration of AI into English language teaching has gained considerable momentum. Teachers are increasingly adopting AI tools to enhance instructional quality and efficiency, motivated by several empirically supported factors.

2.1 AI aids teaching planning

AI tools significantly support teachers in instructional design, saving time and enabling more personalized teaching strategies. In a survey of 156 English instructors conducted by Kohnke and Zou (2025), 78% reported using ChatGPT for lesson preparation. These educators noted that AI helps rapidly generate lesson outlines, exercises, and teaching materials—tasks that traditionally demand substantial effort. By automating routine activities, AI allows teachers to dedicate more time to individualized student interaction and feedback.

AI also streamlines the assessment process. For instance, tools capable of automated grading offer instant feedback on grammar, vocabulary, and short writing assignments (Li et al., 2024). This efficiency reduces the grading burden and enables teachers to concentrate on higher-order instructional goals such as guiding discussions, addressing learning gaps, and designing communicative tasks.

A relevant application lies in writing instruction. Topuz et al. (2025) compared AI-generated scores and feedback with those from human raters using EFL writing samples across proficiency levels. Results indicated that AI effectively identifies surface-level errors and

accelerates feedback delivery, thus freeing teachers from repetitive correction. However, the study also highlighted AI's limitations in evaluating deeper aspects of writing, such as authenticity of argumentation and emotional resonance, which remain areas where human assessment excels.

Moreover, AI facilitates personalized learning. Wang and Chen (2023) observed that AI systems can tailor exercises to individual student levels, assigning advanced vocabulary tasks to proficient learners while reinforcing grammar basics for beginners. Such differentiation fosters engagement and supports progressive skill development.

Despite these benefits, over-reliance on AI poses notable risks. Zhao (2024) cautioned that AI's emphasis on formal accuracy (e.g., grammar and vocabulary) often comes at the expense of communicative effectiveness and real-language applicability. Additionally, reduced peer interaction and feedback opportunities may inhibit the development of critical thinking and social learning (Li et al., 2024). These findings underscore the need for balanced integration—where AI complements, rather than replaces, professional teaching expertise and human interaction.

2.2 Engagement with AI tools in English classes

A significant advantage of artificial intelligence in high school English instruction is its ability to foster student involvement through interactive and gamified learning experiences. By incorporating elements such as timed grammar challenges and interactive story-building tasks, AI enhances teen learners' participation and sustains their motivation—a finding strongly supported by empirical research.

Sun & Kim (2024) conducted a comparative study involving 10th to 12th graders, examining the influence of AI on engagement. The researchers assessed usage frequency through daily tool logs, administered interest surveys, and evaluated writing and speaking skills after each instructional unit. Their results demonstrated that students using AI showed higher consistency in task completion and overall participation. A major contributing factor was the immediate, personalized feedback provided by AI systems, such as praise for coherent arguments or corrective tips for grammatical errors, which heightened students' sense of accomplishment compared to traditional worksheet-based exercises.

Furthermore, AI supports diverse learning styles, enhancing relevance and inclusivity. Park et al. (2023) noted that visual learners benefit from AI-generated concept maps and video examples of real-language use, while auditory learners engage more deeply with spoken dialogue simulations and model speeches, particularly those set in academic or practical scenarios. To evaluate these adaptations, the team employed learning style inventories, classroom observations, and semi-structured interviews. Findings confirmed that tailored AI content increased voluntary participation, encouraged repeated practice, such as rehearsing speeches with AI feedback—and promoted self-reflective learning.

However, sustaining engagement requires balanced implementation. Over-reliance on AI's novelty may lead to "diminishing returns," where initial excitement declines with prolonged

exposure, ultimately reducing motivational benefits. Educators mitigate this risk by blending AI activities with collaborative discussions and human-led instruction, thus preserving social dynamics and supporting long-term engagement.

The strategic incorporation of AI by teachers considerably influences students' learning pathways. Understanding these effects offers critical insights into how AI is reshaping English language acquisition and competency development.

3. How AI Tools Affect Students' Learning Development and Outcomes in English Classes

Artificial intelligence tools are rapidly reshaping the English language classroom, introducing innovative methods to enhance student learning. This section synthesizes recent empirical studies to examine the beneficial impacts of AI, from strengthening oral and written proficiency to fostering cognitive development and enabling personalized instruction.

3.1 How AI supports listening practice

AI-powered listening tools adapt dynamically to students' ability levels. For beginners, audio can be slowed down to 70% of its original speed to aid in distinguishing basic pronunciation and common phrases. Advanced learners, conversely, can challenge themselves by accelerating playback to 120% to train comprehension at a natural or faster-than-normal pace. These systems often incorporate immediate comprehension checks through interactive exercises. For example, after a short dialogue is played, the AI may pose a content-based question such as, "What time will the meeting start?" If the student answers incorrectly, the tool replays the specific segment containing the information—e.g., "We'll have the meeting at 3 PM"—while simultaneously offering vocabulary support, such as highlighting and explaining key terms like "meeting." This iterative process supports autonomous learning and reinforces both decoding skills and lexical knowledge. Yang et al. found that students using these AI listening tools improved their listening skills by 18% over 12 weeks. This was 10% higher than the group using traditional materials. The progress comes from AI's ability to provide clear, adjustable audio, which helps learners focus better and remember specific details.

3.2 AI and Personalized Learning Paths

A major strength of AI is creating personalized learning paths to meet individual needs. Kohnke and Zou found that when teachers use AI in English classes, student engagement increases by 25%. For example, if a student has trouble with past tense, the AI identifies this and suggests extra exercises, like filling in a blank as "She ____ (walked) to school yesterday". If another student struggles with long texts, the AI splits passages into 3 sentence sections and asks questions after each. Teachers observed that students stay focused longer because "the assignments feel specifically tailored to them". Indeed, Kohnke and Zou's research underscores that AI-driven engagement correlates with a marked improvement in learning outcomes—final scores rose by 14% compared to traditional standardized teaching models. This highlights the potential of AI to elevate achievement through personalized and interactive methodologies.

Furthermore, AI tools contribute substantially to English as a Foreign Language (EFL) learning. They mitigate anxiety in oral practice, deliver precise feedback and efficient evaluation in writing, bolster cognitive capabilities such as summarization, facilitate vocabulary acquisition, enhance listening comprehension, and support tailored learning experiences. With careful implementation, educators can leverage these tools to design more effective and inclusive classrooms that promote success for every learner.

That said, these benefits must be weighed against the risks of uncritical adoption. Overuse of AI without strategic oversight can lead to several concerning outcomes: erosion of real-world communication competence, diminished autonomous learning capacity, perpetuation of errors through over-reliance on imperfect AI suggestions, constraint on higher-order cognitive development, and a decline in meaningful social interaction within educational settings. Therefore, while AI undeniably serves as a powerful aid across language domains, including speaking, writing, summarizing, and vocabulary building—its integration demands deliberate instructional design. Without such care, the potential drawbacks pose significant threats to the development of holistic, socially embedded, and critically capable language users.

4. Challenges of Excessive AI Use in English Language Education

Excessive reliance on artificial intelligence in English learning presents considerable risks to students' autonomous learning capabilities and linguistic development, primarily by eroding independent thinking and revealing AI's inherent limitations in processing nuanced language and emotional content.

Foremost among these concerns is how AI overuse diminishes students' capacity for independent learning. Daud et al.'s systematic review emphasizes that heavy dependence on AI leads learners to prioritize automated outputs over active knowledge construction, ultimately weakening their intellectual agency. This finding receives substantial support from Gökoğlu et al. (2025), whose meta-analysis of 120 studies across 28 countries demonstrated that over-reliance on generative AI particularly in grammar practice discourages critical evaluation of AI suggestions and reduces self directed learning efforts. Parallel findings by Hu et al., based on 58 studies from 17 countries, confirmed that excessive AI use hinders original thinking as students increasingly delegate problem solving to automated systems. Lai further contends that such dependency fosters superficial learning behaviors, with students accepting AI generated responses rather than deeply engaging with language concepts. This concerning pattern manifests clearly in specific skill domains. Pei et al.'s research on ChatGPT integration in secondary EFL writing instruction (grades 7 through 12) employed pre tests and post tests, learner surveys, and revision process analysis to demonstrate that uncritical adoption of AI generated content compromises students' writing autonomy by reducing their engagement in drafting and revision processes.

Notably, these detrimental effects are not limited to writing. Emerging evidence indicates that parallel concerns exist in the development of oral proficiency, where over-reliance on AI speaking tools may constrain opportunities for authentic communication and self correction.

Yang et al., who studied EFL university students using peer intelligent agent strategies, employed oral proficiency tests, cognitive load surveys, and anxiety scales to find that excessive reliance on AI agents reduces students' motivation to self-correct. Instead of engaging critically with their speech, learners defer to AI-guided conversations, stunting the development of autonomous oral communication abilities. Collectively, these findings indicate that unchecked AI use across writing, speaking and other language domains risks undermining core objectives of EFL education that include independent thinking, critical evaluation, and self-directed learning.

Second, AI has trouble understanding complex language situations and subtle emotions, which limits how well it helps. Lampropoulos et al. analyzed social media data and found that teachers often notice AI cannot pick up on emotional hints that are really important for getting how real-life conversations work. Lai also says that generative AI struggles with prompts that depend on context, especially those needing cultural or situational knowledge, and this leads to unclear or messy responses.

When it comes to grading writing, Topuz et al. compared how humans and AI score essays. They saw that AI misses the small emotional and contextual details in EFL writing, which means it does not accurately judge how well the writing communicates. In speaking practice, Yang et al. noticed that AI "conversation partners" don't respond well to students' emotional signals. This gets in the way of students developing the practical communication skills they need for real-life interactions. Thus, less independent thinking and AI's struggles with context and emotions clearly show the dangers of trusting AI without question. Addressing these challenges requires intentional strategies to harness AI's potential while mitigating its drawbacks—a balance that hinges on collaborative efforts between teachers and students.

5. How can teachers and students maximize the benefits and minimize the negative effects of using AI in English classes?

To ensure AI serves as a supplementary tool rather than a replacement, both teachers and students should adopt reasoned and collaborative strategies. Teachers, in particular, play a crucial guiding role in the effective integration of AI. They can begin by carefully selecting tools that align with learning objectives. As Kohnke and Zou suggest, platforms like ChatGPT can be used to design engaging activities, such as generating role-playing dialogues that help students practice speaking in meaningful contexts. But teachers should not just rely on what AI generates, they need to check if the content fits their students' level and adjust it if needed. For example, if AI makes a dialogue that's too tough for students, a teacher can simplify the vocabulary to match what the class is currently learning. This way, AI supports the lesson instead of taking over. In writing classes, this approach works too, teachers can use AI as a helper rather than a replacement for their own guidance. Teachers and students can optimize the use of AI in English classes by adopting a balanced and guided approach. For example, Pei et al. suggest allowing students to use AI for initial support such as grammar checks or idea generation, while teachers subsequently provide personalized feedback focused on originality and critical thinking. In a practical scenario, a student might draft an essay about their favorite book using AI for language

corrections, after which the teacher could pose reflective questions like, “Why did you choose this book?” or “What personal connection do you have with the story?” This method helps preserve the student’s unique voice and prevents passive copying of AI content, encouraging deeper reflection and ownership of their writing.

Furthermore, teachers can use AI to stimulate discussion and debate in ways that strengthen critical thinking. Ma et al. (2025) demonstrate that AI generated prompts, such as “Should students have homework on weekends?” can significantly boost student participation and enhance the quality of ideas presented. When students interact with AI in structured formats, for example, through role playing tasks like “argue as an environmental scientist”—they are pushed to evaluate, synthesize, and articulate complex arguments. This reduces the tendency to passively accept AI output and promotes active cognitive engagement.

However, the successful application of these tools depends on active teacher involvement. Instructors should deepen discussions with follow up questions such as, “Can you provide specific examples to support your claim?” In group projects, students might use AI to brainstorm initial ideas, for instance, generating arguments about the importance of recycling and then work together to evaluate, select, and expand upon the most compelling points. This approach allows AI to serve as a creative starting point, while students practice collaboration and develop their own perspectives.

Ma et al. also caution that without thoughtful guidance and structured tasks, students may uncritically adopt AI generated content, leading to superficial learning and over-reliance. Therefore, a balanced integration of AI, with teachers playing a central role in designing and facilitating activities, is essential for maximizing benefits and minimizing risks in the English classroom.

Conclusion

In summary, AI tools have become integral to high school English education, offering significant benefits while also introducing notable challenges.

On the one hand, AI enhances learning in multiple dimensions: it reduces anxiety in speaking practice, provides immediate feedback to improve writing, adapts listening exercises to individual proficiency levels, and enables personalized learning experiences. For teachers, AI streamlines lesson planning and automated grading, freeing up time for more meaningful instructional interaction and individualized support. These advantages underscore the potential of AI as a valuable supplement to traditional English language instruction.

On the other hand, over-dependence on AI poses serious risks. It can diminish students’ motivation to think independently, encourage uncritical adoption of generated content, and undermine the development of original problem-solving skills. Furthermore, AI’s current limitations in interpreting emotional subtleties and cultural context reduce its effectiveness in fostering genuine communicative competence. Such shortcomings highlight the irreplaceable role of human interaction in language learning.

The impact of AI on student learning is thus dualistic: it enhances efficiency and skill

development, but also threatens to promote passive learning behaviors. Navigating this tension requires deliberate collaboration between teachers and students. Educators should critically select and adapt AI tools to support meaningful learning experiences, while guiding students toward reflective and critical use of AI. Learners should engage actively with AI outputs by questioning suggestions, refining AI-generated responses, and committing to the intellectual effort required for deep learning.

Ultimately, AI serves as a powerful aid rather than a comprehensive solution. With balanced implementation and mindful usage, it can help high school students strengthen not only their English language skills, but also their capacity for autonomous thinking and critical engagement.

Works Cited

- 1.Daud, A., Aulia, A. F., Muryanti, Harfal, Z., Nabilla, O., & Ali, H. S. (2025). Integrating artificial intelligence into English language teaching: A systematic review. *European Journal of Educational Research*, 14(2), 677–691.
- 2.Huang, A. Y. Q., Lin, C.-C., Su, S.-Y., Yen, R.-X., & Yang, S. J. H. (2025). Improving students' learning performance and summarization ability through a Generative AI-enabled Chatbot. *Educational Technology & Society*, 28(3), 82–111.
- 3.Hu, D.-X., Pang, D.-D., & Xing, Z. (2025). Evaluating the effects of Generative AI on student learning outcomes: Insights from a meta-analysis. *Educational Technology & Society*, 28(3), 226–240.
- 4.Topuz, A. C., Yıldız, M., Taşlibeyaz, E., Polat, H., & Kurşun, E. (2025). Is generative AI ready to replace human raters in scoring EFL writing? Comparison of human and automated essay evaluation. *Educational Technology & Society*, 28(3), 36–50.
- 5.Yang, G., Rong, Y., Zou, Z., Fang, X., Yang, M., & Zeng, Q. (2025). An empirical study of the effects of peer intelligent agent instructional strategies on EFL university students' oral English learning performance, cognitive load, learning anxiety and technology acceptance. *Educational Technology & Society*, 28(3), 318–336.
- 6.Lai, C.-L. (2025). Consolidate knowledge or build scientific models? The role of online information-searching strategies in students' prompt sequences with Generative Artificial Intelligence. *Educational Technology & Society*, 28(3), 241–262.
- 7.Kohnke, L., & Zou, D. (2025). The role of ChatGPT in enhancing English teaching: A paradigm shift in lesson planning and instructional practices. *Educational Technology & Society*, 28(3), 4–20.
- 8.Lampropoulos, G., Ferdig, R., & Kaplan-Rakowski, R. (2025). A social media data analysis of general and educational use of ChatGPT: Understanding emotional educators through Twitter data. *Educational Technology & Society*, 28(3), 51–6514.
- 9.Pei, L., Jong, M. S.-Y., Huang, B., Pang, W.-C., & Shang, J. (2025). Formally integrating Generative AI into secondary education: Application of ChatGPT in EFL writing instruction. *Educational Technology & Society*, 28(3), 281–297.
- 10.Yang, G., Rong, Y., Zou, Z., Fang, X., Yang, M., & Zeng, Q. (2025). An empirical study of the effects of peer intelligent agent instructional strategies on EFL university students' oral English learning performance, cognitive load, learning anxiety and technology acceptance. *Educational Technology & Society*, 28(3), 318–336.
- 11.Zhao, X. (2024). Research on the application of AI in English teaching after - class practice. *Journal of Educational Research*, 15(3), 25 - 32.
- 12.Ma, Z.-Q., Cui, X., Liu, W., Tu, Y.-F., & Hwang, G.-J. (2025). ChatGPT-assisted collaborative argumentation: Effects of role-playing prompts on students' argumentation outcomes, processes, and perceptions. *Educational Technology & Society*, 28(3), 133-150.
- 13.Hakimi, L., Eynon, R., & Murphy, V.A.(2021). The ethics of using digital trace data in education: A thematic review of the research landscape. *Review of Educational Research*, 91(5), 671-717.
- 14.Chen, M.-R. A. (2024). Metacognitive mastery: Transformative learning in EFL through a generative AI chatbot fueled by metalinguistic guidance. *Educational Technology & Society*, 27(3), 407-427.

Comparative Study of Parametric and Non-Parametric Models in Crop Recommendation

By Krishna Ramakrishnan

Abstract

Agriculture is facing mounting challenges from population growth, food insecurity, soil degradation, and climate variability, making traditional practices insufficient to sustain production. AI and machine learning (ML) have recently been applied to agriculture, offering opportunities to improve crop selection and land use through data-driven recommendations. Previous studies have explored crop recommendation systems, but their findings have often been inconsistent due to differences in datasets and methods. This study addresses that gap by benchmarking parametric and non-parametric models on the same dataset to evaluate their performance in a multi-class crop recommendation task. Using an agricultural dataset of 8,000 entries across 11 crop types, data preprocessing included one-hot encoding, label encoding, and median imputation. A neural network implemented in TensorFlow was compared against K-Nearest Neighbors (KNN) and Random Forest models. Results showed that the neural network achieved 80–83% accuracy, while the non-parametric models remained near 9–10%, close to random guessing. These findings suggest that parametric models are better suited for capturing the complex, non-linear patterns in agricultural data. Despite limitations such as overfitting and reliance on a single dataset, the study highlights the potential of AI to provide more accurate and sustainable crop recommendations, ultimately supporting global food security.

Introduction

Agriculture has remained one of the most essential sectors for sustaining human life, yet it has faced increasing challenges in recent decades. With the global population continuing to rise and food insecurity becoming a growing issue, farmers must meet higher production quotas while dealing with environmental pressures, soil degradation, and unpredictable climate conditions (Shields 2024). Traditional farming practices have been proven insufficient to address these complex issues, even if they are the foundation of agriculture. In recent years, technological innovations have emerged to support agriculture, including advanced sensors for monitoring soil and crop health, innovations in crop engineering, and software systems for optimizing production. These advancements have started a shift toward smart agriculture, where data-driven decisions have improved efficiency, reduced waste, and optimized yields (Shields 2024). Such methods have been crucial in supporting sustainable food production and meeting the demands of a rapidly changing world.

In the current state of our world, populations are constantly growing, food insecurity is rising and the agricultural sector is struggling (Shields 2024). Several fields, including finance, healthcare, e-commerce, and many more have been revolutionized by AI. Farmers throughout the world are struggling due to problems such as selecting the best crops for their soil and climate, optimizing land use, and adapting to changing local conditions (Waikar et al., n.d.). As a result, the implementation of AI and data science has become crucial to increase efficiency and meet global food demand. AI and ML-based crop recommendation systems have the potential to

address these issues by analyzing large datasets to generate accurate and even location-specific recommendations. While prior studies have used machine learning to create crop recommendation systems, their results have been scattered and inconsistent (Waikar et al., n.d.). This study aimed to address this gap by comparing the performance of parametric models and non-parametric models on a uniform dataset. The specific objective was to observe differences in performance between a neural network and two non-parametric models (K-Nearest Neighbors and Random Forest) to evaluate their suitability for a multi-class crop recommendation task.

Materials and methods

The dataset used for this project was publicly available on Kaggle and is licensed under MIT (Shankar 2025). It contains distinct soil and environmental input features including temperature, humidity, moisture, soil type, nitrogen, phosphorus, potassium, and fertilizer used. There are 8000 data points in this dataset with 11 different crop classes.

Data preprocessing followed conventional standards in the field. The dataset was loaded using the Pandas library, and the input features were separated from the target variable (crop type). All categorical features were converted into numerical format using one-hot encoding, while the target variable was converted into integer labels using scikit-learn's LabelEncoder. Any missing values were filled with the median of their respective columns. The dataset was split into training and testing sets in a 70:30 ratio.

A neural network was implemented using Keras, the API for TensorFlow. The model consisted of three hidden layers, each using the ReLU activation function. Batch normalization was applied between layers to improve training speed (Kumar, 2024). The output layer used a softmax activation function to classify the data into 11 crop types. The model was compiled with the Adam optimizer and a Sparse Categorical Cross-Entropy loss function. Training was done over 2,500 epochs with a batch size of 64 and a validation split of 20%. Model performance was measured using accuracy, calculated by comparing predicted class labels with the true labels in the testing dataset.

The K-Nearest Neighbors model was implemented using scikit-learn's KNeighborsClassifier, with an initial hyperparameter of 8 neighbors (scikit-learn, 2019). A hyperparameter optimization loop was created to test values of k from 1 to 100, recording accuracy for each. The Random Forest model was implemented using scikit-learn's RandomForestClassifier with 91 trees as the main hyperparameter. Similar to KNN, an optimization loop was used to find the number of trees with the best accuracy. Both models were trained on the training dataset and tested on the testing dataset. Accuracy was calculated by comparing predicted labels to actual labels. To demonstrate model predictions, a single sample from the test set was classified and the result was converted back to its corresponding crop name.

Results

The dataset used this study consists of 8000 records with a mix of numerical and categorical features important to agricultural conditions, including Temperature, Humidity, Moisture, Nitrogen, Phosphorous, Potassium, Soil Type, and Fertilizer Name, with Crop Type as

the target variable (Shankar 2025). To keep the data clean and usable, all numerical columns were checked and converted to the correct format, and any missing values were filled in using the median value of each column. Categorical columns were encoded using one-hot encoding for model training and label encoding for visualizations.

A pair plot of all features (Figure 1) revealed heavy overlap among crop classes, especially in features related to soil nutrients and environmental conditions. There were no clear linear separations between classes, showing the need for complex, non-linear models to effectively capture hidden patterns. The correlation matrix (Figure 2) confirmed that most features are weakly correlated with each other, with the exceptions of a moderate positive correlation between Temperature and Humidity ($r = 0.53$), and a strong negative correlation between Nitrogen and Phosphorus ($r = -0.64$). Because every feature adds something unique, all 8 can be kept without worrying about redundant information.

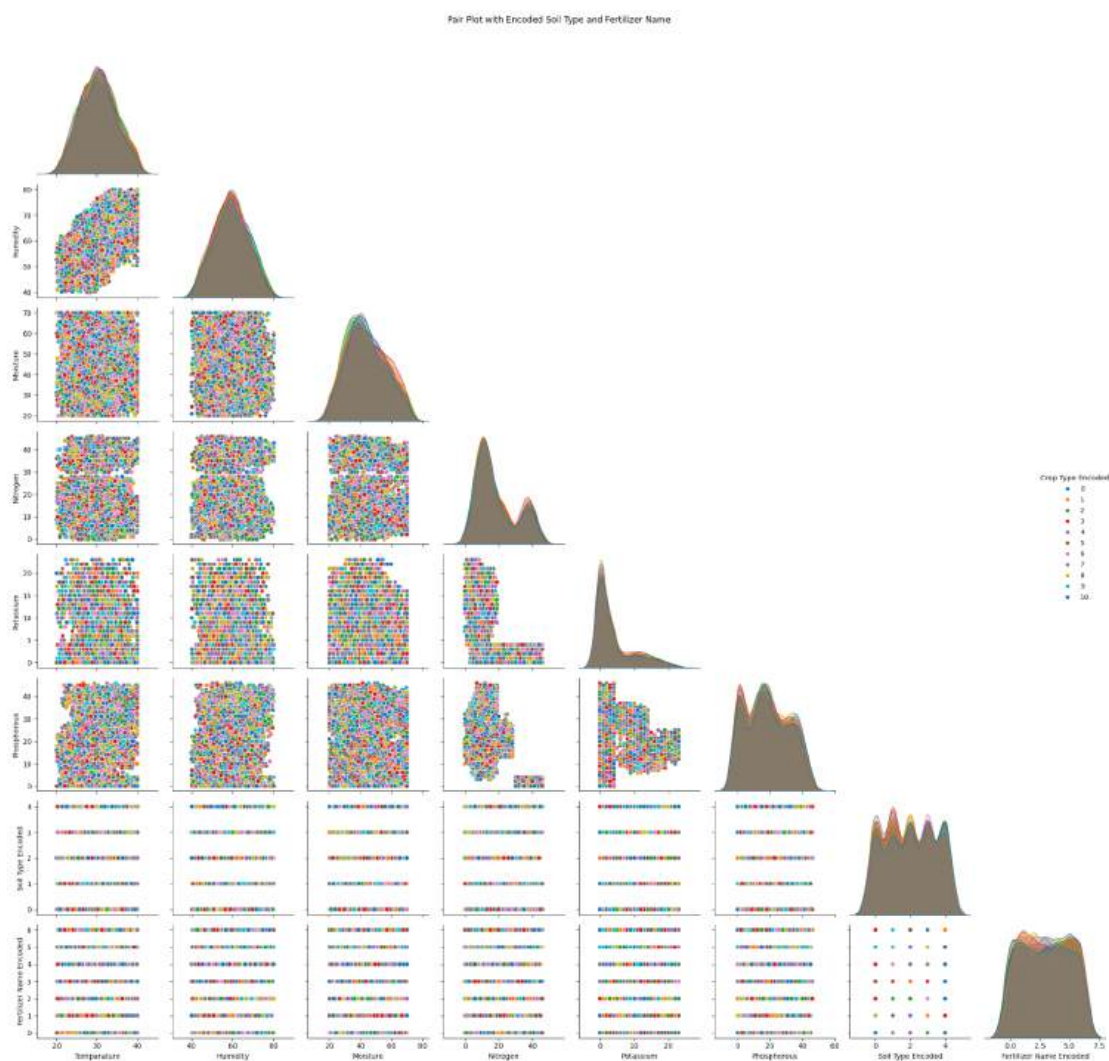


Figure 1: This pair plot shows the relationships between different factors like temperature, humidity, and nutrients in the soil. The plots along the diagonal show how each individual factor is spread out. The other plots show how two different factors relate to each other. The different colors represent various crop types, which helps to see how these relationships change depending on the crop

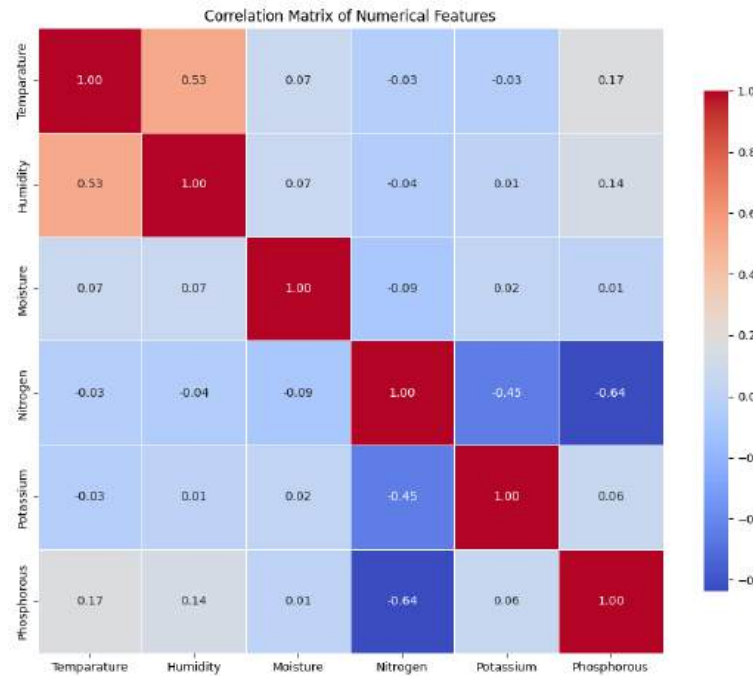


Figure 2: This matrix shows how the different features relate to each other. Temperature and Humidity have a strong positive relationship, meaning they tend to increase or decrease together. Nitrogen has a negative relationship with Phosphorus and Potassium, suggesting that when nitrogen levels are high, the others tend to be low.

The neural network achieved an accuracy score of 80% around the 2000th epoch and fluctuated between 80 to 83% for the remaining epochs (Figure 3). However, the validation accuracy consistently stayed at 9 to 11%, indicating that there was significant overfitting of the data (about the same accuracy as random guessing). Loss continued to decrease until the model converged. It was able to classify the majority of the crop types but failed in cases where 2 crop types had very similar ideal conditions.

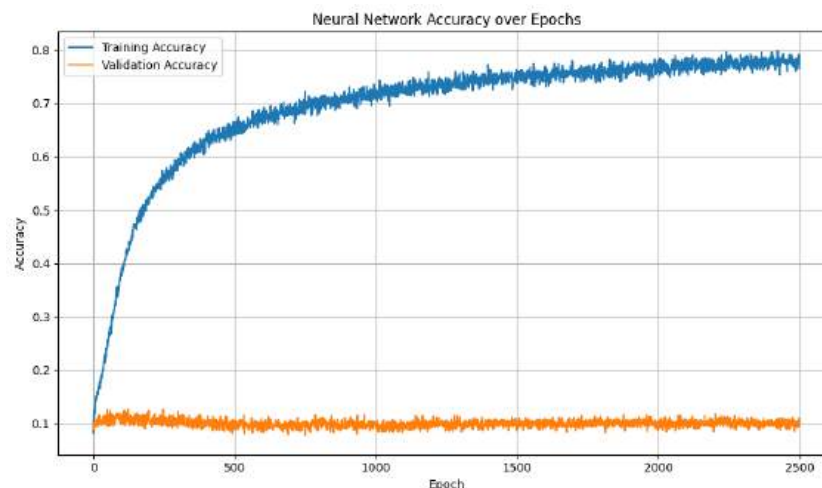


Figure 3: Neural network training and validation accuracy over 2500 epochs. The neural network with 3 hidden layers was trained on soil and environmental data (8000 data points, 8 features, 11 crop types in the target variable). Training accuracy (blue) increased steadily over time until it plateaued at 80-83% accuracy after the 2000th epoch. Validation accuracy (orange) remained near 10%, indicating overfitting.

The KNN model achieved a max accuracy score of 10.2% (at $k = 8$) across all tested values of k from 1-100. The lowest accuracy score of 8.58% occurs at $k = 5$. The regression line in the graph between k values and accuracy (Figure 4) shows that there is almost no correlation between the 2 (correlation coefficient of 0.11).

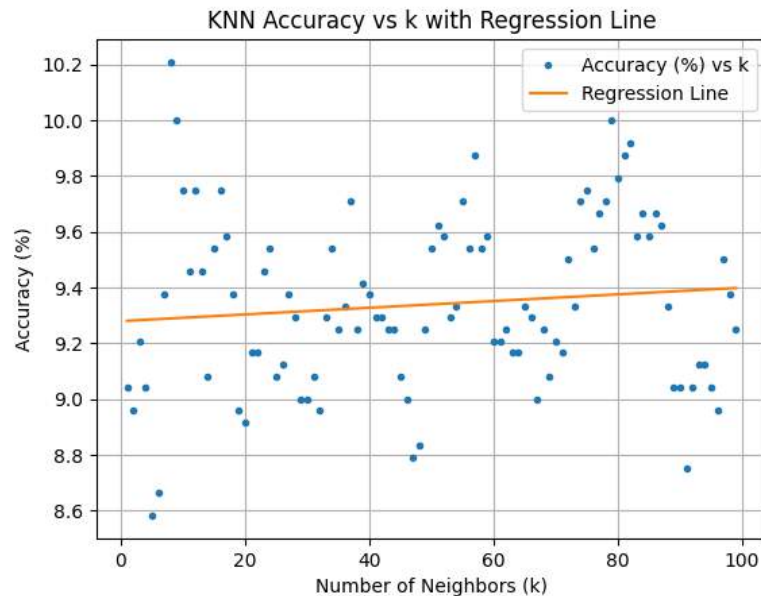


Figure 4: This graph shows the model's accuracy as we increased the number of trees it used. The accuracy went up and down, peaking around 9.4%. After about 100 trees, the accuracy became more stable, but didn't consistently improve.

Discussion

This study explored the role of artificial intelligence and machine learning in agriculture, focusing on the development of a crop recommendation system. Agriculture today faces immense challenges such as soil degradation, unpredictable climate, and the need for higher yields to meet food demands. AI-driven approaches enable the processing of large datasets and provide precise crop recommendations that traditional methods cannot achieve. To evaluate this potential, three models were selected for comparison: a neural network representing parametric methods, and K-Nearest Neighbors (KNN) and Random Forest representing non-parametric methods. Using a standardized dataset of 8,000 entries containing soil and environmental features, each model was trained and tested on the same split of data.

By comparing the parametric approach (neural network) with the non-parametric models (KNN and Random Forest), we directly compared different machine learning models on a multi-class crop classification task in agriculture. The neural network achieved 80% accuracy around the 2000th epoch and maintained a range of 80% to 83% for the remainder of the training (Figure 1). This indicated good generalization and model stability. On the other hand, the non-parametric models performed poorly. The KNN model, with a hyperparameter of 70 neighbors, achieved an accuracy of only 9.21% (Figure 4), and testing various values of k from 1 to 100 did not improve the results. Similarly, the Random Forest classifier, using 100 trees, also

achieved an accuracy of just 9%, showing limited promise (Figure 5). These results suggest that the parametric neural network may be a better option for capturing the hidden patterns and complexities in the dataset, while the non-parametric models struggled due to high dimensionality in the data set, or due to difficulties in the multiclass classification task (Hardesty 2017). Overall, the findings support that parametric models such as neural networks perform better in a multiclass application such as a crop recommendation system.

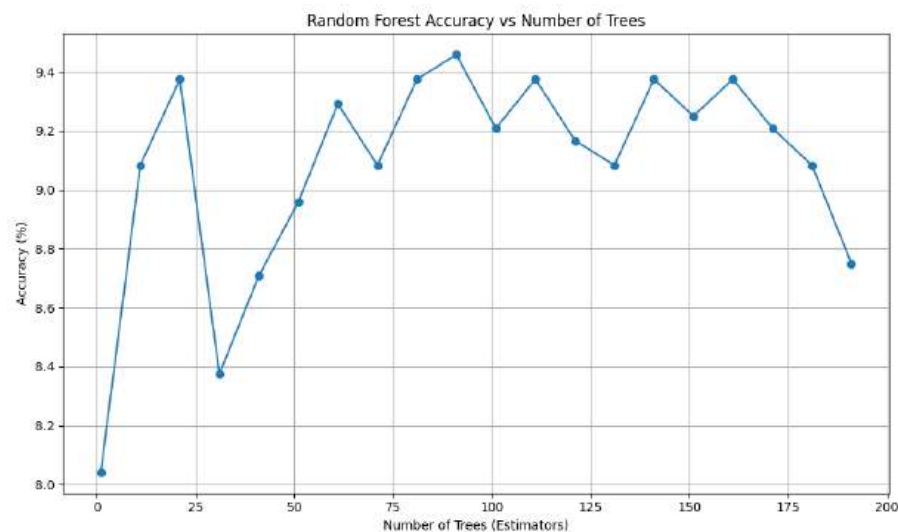


Figure 5: This graph shows the model's accuracy as we increased the number of trees it used. The accuracy went up and down, peaking around 9.4%. After about 100 trees, the accuracy became more stable, but didn't consistently improve.

These findings indicate that parametric models such as neural networks are better suited for multi-class agricultural recommendation systems, as they can capture non-linear patterns in high-dimensional data. However, the neural network's overfitting suggests that more thorough preprocessing, extra techniques to reduce overfitting and improve generalization, or access to larger and more diverse datasets would be necessary to improve generalization. If more time and resources were available, the study could be extended by testing other advanced architectures such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), or transformers, and by including real-world field data beyond the Kaggle dataset. Limitations of this study include reliance on a single dataset, limited hyperparameter exploration, and a lack of evaluation using domain-specific metrics such as crop yield efficiency or economic viability. Despite these constraints, the work highlights the potential of AI in agriculture. If fully fleshed out, an accurate crop recommendation system could significantly reduce resource waste, improve food security, and provide small-scale farmers with personalized insights that directly impact their way of life and contribute to sustainable agricultural practices worldwide.

Works Cited

- Shankar. (2025, January 28). Crop and soil dataset. Kaggle.
<https://www.kaggle.com/datasets/shankarpriya2913/crop-and-soil-dataset>
- Waikar, V., Thorat, S., Ghute, A., Rajput, P., & Shinde, M. (n.d.). Crop Prediction based on Soil Classification using Machine Learning with Classifier Ensembling. In International Research Journal of Engineering and Technology. Retrieved August 24, 2025, from <https://www.irjet.net/archives/V7/i5/IRJET-V7I5931.pdf>
- scikit-learn. (2019). sklearn.neighbors.KNeighborsClassifier — scikit-learn 0.22.1 documentation. Scikit-Learn.org.<https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>
- Hardesty, L. (2017, April 14). Explained: Neural networks. MIT News; Massachusetts Institute of Technology. <https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414>
- Kumar, R. (2024, May 20). Batch Normalization: Theory and TensorFlow Implementation. Datacamp.com; DataCamp.
<https://www.datacamp.com/tutorial/batch-normalization-tensorflow>
- Shields, R. (2024). The Use of Ai (Artificial Intelligence) in Agriculture & Farming. Wwww.agrirs.co.uk.
<https://www.agrirs.co.uk/blog/2024/02/the-use-of-ai-artificial-intelligence-in-agriculture-and-farming?source=google.com>

Other Tables & Figures

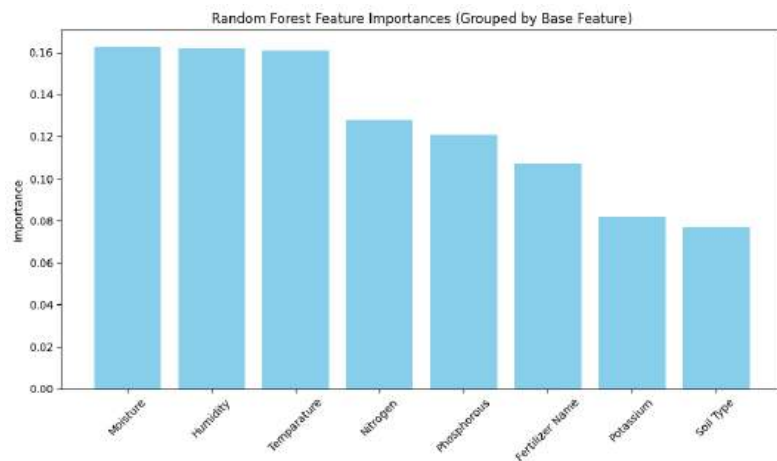


Figure 2: This chart shows which features were most important for the Random Forest model's predictions. The model relied most heavily on Moisture, Humidity, and Temperature, while Potassium and Soil Type were the least important.

A Comprehensive Review of Depression: Risk Factors, Anatomy, and Treatment

By Tanvi Shrowty

Abstract

Major depressive disorder (MDD) is characterized by a depressed mood and anhedonia and has a lifetime prevalence of around 10-15% of the global population. Given the wide array of people who suffer from this disorder, there is an extensive amount of research done investigating its biological causes and therapy for successful depression treatment. Despite this research, current therapeutic approaches may be flawed because of the trial-and-error and “one-size-fits-all” strategies that are being used. The underlying causes and factors for each individual with depression differ greatly, so the same treatment may not be successful for all. This paper seeks to describe the risk factors and biological elements of depression in an attempt to better understand the future of individualized care in depression. Variations in brain circuitry and anatomy are often found in patients with depression. These variations include changes in the prefrontal cortex, amygdala, and hippocampus, as well as varying levels of neurotransmitters such as dopamine and serotonin. Additionally, there are varying treatment methods available for depressed patients. Treatments can be separated into different categories. There are pharmaceutical drugs that target dopamine or serotonin as well as various types of talk therapy. Given the differing underlying factors contributing to depression, it is necessary to further research the development of individualized treatment for each patient instead of the repetitive approaches that are currently utilized.

Introduction

Major depressive disorder (MDD) can be defined as a mental illness that negatively impacts an individual’s emotions, thoughts, actions, and perceptions of the world (American Psychiatric Association, 2013). The symptoms of depression range from mild to severe, but any presence of this mental illness negatively impacts the lives and wellbeing of those diagnosed with depression.

Depression is diagnosed through clinical interviews and symptom assessments (Schmidt et al., 2011). Clinicians look for the symptoms of depression, such as evaluating mood or sleep patterns, to diagnose a person with the mental illness. To be diagnosed with depression, clinicians refer to the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition Text Revision (DSM-5 TR) (American Psychiatric Association, 2013). The patient must have five or more symptoms present within a two week period for it to classify as a Major Depressive Episode. At least one of the symptoms must be a depressed mood or anhedonia (inability to feel pleasure) (Tolentino & Schmidt, 2018). Other symptoms essential to the diagnosis include: changes in appetite or weight, insomnia or hypersomnia, psychomotor agitation or retardation, fatigue, difficulty thinking or concentrating, feeling worthless or guilty, and suicidal thoughts. This method of diagnosis is only based on the clinical evaluation of depressive symptoms and does not consider the biological differences behind them. This provides the opportunity for two

patients with the same symptoms to be given similar, if not identical, treatment, even though there are different biological factors that give rise to these symptoms.

Depression is one of the most commonly diagnosed mental disorders, with an estimated lifetime prevalence of 10-15% (Berton & Nestler, 2006). Certain individuals are more likely to experience depression than others because of their genetic factors and environmental context. For example, females, ages 18-24, low education, low income, and unmarried individuals have higher incidences of depression (Kessler et al., 2005). It is important to note that being a part of these demographics is not a direct cause of depression, but rather, they are correlated with being at high risk for the disorder.

There are many methods of treatment for depression, but the focus of this paper is two categories: pharmaceutical drugs and talk therapy. Two of the primary neurotransmitters associated with depression are the monoamine neurotransmitters dopamine and serotonin. Specifically, dopamine is considered both a monoamine and a catecholamine, while serotonin is just a monoamine. In many cases, antidepressants target dopamine in depressed brains. Some drugs, such as phenelzine, tranylcypromine, and isocarboxadid, function to increase monoamine levels by inhibiting the enzyme (monoamine oxidase) that breaks down monoamines. These drugs are known as monoamine oxidase inhibitors (MAOIs) (Papakostas, 2006). Other dopaminergic-based drugs, such as dextroamphetamine, pemoline, and methylphenidate, focus on preventing the reuptake of dopamine (Cardenas et al., 2004; Volkow et al., 2002). Additionally, there are serotonergic-based approaches. The most commonly prescribed antidepressants are selective serotonin reuptake inhibitors (SSRIs), which increase serotonergic neurotransmission by blocking the reuptake of serotonin, allowing serotonin to stay active in the brain for longer (Stahl, 1998). An alternative method, or sometimes used in tandem with pharmaceutical drugs, for treating depression is talk therapy. Talk therapy involves working with a certified practitioner to evaluate factors that could contribute to a patient's depression and modify a patient's method of thinking to alleviate their symptoms (Wright & Beck, 1983). There are many different kinds of therapy, including cognitive behavioral therapy and interpersonal therapy.

This paper encompasses a review of the circuitry of MDD (the brain regions and risk factors associated with depression) and treatment methods used for depression. Throughout this paper, the complexity and variation in the brains of each depressed patient will be emphasized, demonstrating the need for investment in promoting individualized care.

Risk Factors

Certain risk factors make a person more likely to develop depression. These risk factors can be traits that an individual is born with (genetic) or results of the situation around them (environmental). Risk factors interact with each other in complex ways, and a combination of both genetic and environmental risk factors may be present in depressed patients. However, risk factors do not cause depression; rather, people who have these qualities are at a higher risk for depression. It is important to understand risk factors because it differentiates depressed patients

from each other and can identify which therapies can be more successful.

Genetic Risk Factors

Family and twin studies provide evidence that the genetic risk factors giving rise to depression can be inherited (Shadrina et al., 2018). In fact, data demonstrates that the risk for depression increases two to three times in the offspring of a parent with depression (Sullivan et al., 2000). However, depression is polygenic, meaning that depression is not defined by a single gene in an individual. Genetic forms of depression can result from many genes, and may additionally arise due to compounding environmental factors (Ebmeier et al., 2006).

The dysfunction of dopamine and serotonin neurotransmission are associated with depression in individuals. Many studies demonstrate how decreased dopaminergic and serotonergic activity are underlying factors of depression (Zhao et al., 2022). Genes give rise to factors controlling dopamine and serotonin levels in the brain, such as the reuptake of these neurotransmitters or the amount of transmitters made. These genes are inherited by offspring from parents, meaning that any variants leading to the genes' dysfunction can be inherited as well.

For example, one specific gene that is found to function differently in depressed brains is *SLC6A4*, which encodes the serotonin reuptake transporter localized to the synaptic cleft to the presynaptic neuron (Shadrina et al., 2018). The activity of *SLC6A4* has an impact on the maintenance of serotonin in the synaptic region. Some individuals possess a *SLC6A4* variant resulting in reduced gene expression, meaning that less serotonin transporters are made in their brain. Fewer serotonin transporters causes less serotonin reuptake, which leads to the altered serotonin levels underlying depression. If patients possess a dysfunctional variant of *SLC6A4*, either resulting from inheritance or a sporadic mutation, it can increase the risk for depression diagnosis.

Another gene that functions differently in depression is *BDNF* (Brain-Derived Neurotrophic Factor). Normally, *BDNF* function is associated with the neuroplasticity and neurogenesis of the adult brain (Shadrina et al., 2018). However, some people possess a variation to this gene that causes impaired processing of the BDNF protein. This results in altered BDNF function, which is an underlying factor of impaired neuroplasticity and both developmental and adult neurogenesis (Chen et al., 2004). Studies have demonstrated an association between depression and neuroplasticity, meaning that possessing the dysfunctional variant of BDNF could give rise to depressive symptoms. Also, BDNF is connected to the function of serotonergic neurons (Shadrina et al., 2018), meaning that BDNF dysfunction may also lead to the dysfunction of serotonin in the brain.

Genetic variants alone do not guarantee a depression diagnosis in all cases. When combined with external environmental risk factors, the probability of diagnosis increases.

Environmental Risk Factors

Environmental factors play a key role in the development of depression. Environmental

risk factors consist of negative external factors or circumstances, such as dysfunctional upbringings or social lives. These factors can and may interact with the genetic differences in the brain to possibly increase the likeliness of depression in an individual (Hammen, 2018).

A common risk factor for depression are stressors or stressful circumstances. Studies demonstrate that over 80% of people with clinically significant depression have experienced or are experiencing severely stressful events, such as abuse or financial insecurity (Brown & Harris, 1989). Chronic stress is correlated with overactivity in the hypothalamic-pituitary-adrenal (HPA) axis, which is associated with structural changes in the hippocampus. These changes in the hippocampus can give rise to depressive symptoms in a person (Lei et al., 2025). Exposure to stressful conditions in childhood, such as child abuse or parental conflict, can be manifested as depression throughout an individual's life since early exposure to stressors lowers the amount of stress needed to trigger a major depressive episode in a patient (Hammen et al., 2000).

In addition, studies demonstrate that women are far more likely to develop depression in their adolescent or adult years. Women possess a 2:1 greater likelihood of experiencing depression than men (Hammen, 2018). Factors that provide the basis for women developing depression include biological elements, such as hormonal changes, or environmental aspects, such as issues with body image in maturing girls (Hammen, 2018). Some evidence suggests that women have a greater exposure to stressors, such as domestic violence or sexual abuse (Barth et al., 2013). This increase in stress in a woman's life has the potential to give rise to the biological changes in the brain that are associated with depression.

There are many different environmental factors that put people at a higher risk for depression. The risk factors that underlie the onset of depression in an individual vary greatly amongst patients, demonstrating the importance of individualized treatment for each person.

Anatomy of Depression

Depression is characterized by numerous psychological and cognitive symptoms, which are rooted in biological systems and dysfunction of the neural network in the brain (Trifu et al., 2020). There are clear neurological differences found in individuals with depression as compared to typically functioning brains. There could be structural differences, such as changes in volume or connectivity. On the other hand, changes in function may be present, such as differences in metabolism or activity. These structural differences in the brain may lead to the symptoms of depression in a patient. Studies consistently implicate three primary brain regions in the anatomical dysfunction of MDD: the prefrontal cortex, amygdala, and hippocampus.

Prefrontal Cortex (PFC)

Studies repeatedly demonstrate that the PFC is implicated in depression (Pandya et al., 2012). The PFC is responsible for planning and directing behavior over time. (Kolb et al., 2012). Additionally, the PFC projects information to other areas, such as the hippocampus and amygdala, which contributes to higher-order processing of executive function (Hathaway & Newton, 2023). The PFC synthesizes information received from the temporal and parietal lobes

to allow complex, goal-directed behavior (Rolls, 2023). The PFC is able to adapt to different situations, making this region essential for survival.

The PFC consists of three major regions that are affected in depressed brains: the dorsolateral prefrontal cortex (dlPFC), the orbital and medial paralimbic cortex, and anterior cingulate cortex (ACC).

The dlPFC is involved in cognitive control, solving complicated tasks, and keeping information and memories. The dlPFC has been shown, through neuroimaging and MRIs, to be altered in depressed brains. Hypoactivity as well as a decreased blood flow and metabolism are present in the dlPFC of patients with depression (Rigucci et al., 2009; Kimbrell et al., 2002). In fact, the left dlPFC is shown to involve human motivation by reward punishment systems (Trifu et al., 2020). Individuals with depression are known to not be motivated by reward systems, which is a symptom that could correlate with the lack of metabolism and blood flow in the area.

The orbital and medial paralimbic cortex corrects and controls the emotional responses that originate from the amygdala. In depression, the orbital and medial paralimbic cortex is shown to have large volume decreases, particularly in the orbitofrontal cortex (OFC) (Palazidou, 2012). Because there is a parallel between the OFC and emotional response, a smaller volume in the OFC may correspond with decreased function and therefore a change in emotional behavior, which is a characteristic found in patients with depression.

Furthermore, the ACC plays a vital role in depression through its association with cognitive responses to emotional stimuli (Vogd, 1992; Critchley, 2004; Etkin, 2006). Additionally, it has many connections to other brain regions, like the amygdala. These connections are altered in depressed brains through a combination of increased and decreased connectivity. These irregularities in connectivity underlie impaired communication between the ACC and other brain regions, which could be associated with the overly negative perceptions that people with depression experience (Wang et al., 2023). Also, in depression, the ACC experiences a large decrease in volume and metabolism (Pandya et al., 2012; Trifu et al., 2020; Drevets et al. 1997), which could alter the functionality of the ACC in patients with depression, leading to emotional irregularity. One part of the ACC (the ventral ACC) is involved with the stress response (Pandya et al., 2012). Stress and depression are closely related. Stress is considered a risk factor for depression and having depression may make an individual feel more stressed. Structural changes in the ACC during depression may underlie stress, because the ACC is a vital part of stress response.

Amygdala

The amygdala is a vital part of systems involved in generating emotion and memory processes (Aggleton, 1992). The amygdala plays an important role in expressing emotions, emotional learning, and is involved in detecting and responding to nonspecific stimuli by sending signals to other parts of the brain (Trifu et al., 2020). Additionally, neuroimaging studies show that the amygdala signal is stronger in response to negative stimuli than positive stimuli (Whalen et al., 2002). The amygdala is a key part of processing fear expression and fear

response. For example, the amygdala has been shown to contribute to distinguishing threatening stimuli (faces) from neutral stimuli (Adolphs et al., 1995; Adolphs et al., 1998).

Structural changes in the amygdala, such as changes in volume, are often observed in depression. Decreased volume in the amygdala has been reported in patients with depression (Sheline et al., 1998). In fact, a neuroimaging study demonstrated that individuals with an early exposure to traumatic events in adolescence had decreased amygdala volumes throughout their lives. These early life stressors are a key risk factor in depression. The volume of the amygdala keeps decreasing as the individuals enter adulthood, demonstrating how vital a role environmental risk factors can play later in life in the development of depression (Korgaonkar et al. 2013).

Furthermore, functional changes within the amygdala have been observed in depressed brains. Hyperactivity in the amygdala has been associated with severity of depression. Indeed, elevated activity is positively correlated with the severity of depression (Palazidou, 2012; Whalen et al., 2002). Abnormally elevated resting cerebral blood flow has also been seen in the amygdalas of depressed brains (Drevets et al., 1992). This hyperactivity in the amygdala has the possibility to disrupt wider brain systems, which underlies symptoms found in depression (Dougherty & Rauch, 1997). On the other hand, the amygdala has also been shown to experience hypoactivity and a slower metabolism in depressed brains. Depressed children have shown a muted amygdala response to fearful stimuli, suggesting that hypoactivity in the amygdala may be related to depression in adolescence (Thomas et al., 2001). The amygdala responses to stimuli have also been shown to differ in depressed patients. For example, there have been reports of the duration of the amygdala response to negative stimuli being prolonged in individuals with depression (Siegle et al., 2002). Also, patients have shown hyperactivity in the amygdala when shown negative stimuli, but at the same time, hypoactivity in response to positive stimuli (Trifu et al., 2020).

The amygdala is a known region for depression, but can be altered in many different ways in the disorder's underlying circuitry, demonstrating the need for different and varied approaches to depression treatment.

Hippocampus

The hippocampus is a vital part of learning and memory in the brain. Changes in hippocampal structure, connectivity, and function have been observed in depression. (Pandya et al., 2012). Additionally, the hippocampus has a key role in emotional and stress response, which it accomplishes in part through connections to other brain regions like the HPA axis (Trifu et al., 2020). The hippocampus supports neurogenesis and neuroplasticity even in adulthood, which is largely important to the brain's ability to adapt to stress, manage emotions, store memories, and learn.

Many MRI studies show volume reduction in the hippocampus in depressed brains, of which some report reductions ranging from 8-19% (Bremner et al., 2000; Sheline et al., 2002). There are some individuals that have been found to have exceptionally small hippocampal

volumes. For example, adolescents who have experienced psychotrauma in early childhood are found to have small hippocampi (MacQueen & Frodi, 2011). Experiencing trauma at a young age is associated with being at a high risk for depression, which could be related to the small hippocampus volume found in these individuals.

Furthermore, volume reduction in the hippocampus can lead to dysfunction, which is associated with behaviors found in depressed individuals. Studies demonstrate association between depression and memory impairment (Burt et al., 1995). The hippocampus plays a vital role in forming and storing memories, meaning that any change to its ordinary structure, such as the volume reduction found in depressed states, may give rise to changes in the memories of people with depression. Also, the hippocampus plays an important part in emotional regulation, which can be impaired due to structural changes. Hippocampal dysfunction is responsible for responses based on emotions and environmental factors rather than logic or memory (Trifu et al., 2020). Links have been found between impulsive behavior and suicidal tendencies in depression (Swann et al., 2008), a pattern that could be related in part to the impaired hippocampal ability after volume reductions in depression.

Clearly, many of the depressive symptoms are attributed to dysfunction in numerous regions of the brain. These different regions converge on similar behaviors, making it necessary to treat the dysfunction in the brain regions to treat the symptoms.

Treatments for Depression

Dopamine Based Approaches

Dopamine is a neurotransmitter and one of the most abundant catecholamines in the brain. Other important catecholamines include norepinephrine and epinephrine. It plays a key role in responses to pleasure and reward, motivated behavior, motor control, attention, and hormone secretion (Zhao et al., 2022). Decreased dopaminergic activity is found in patients with depression. This is attributed to the fact that many of the processes that dopamine contributes to are impaired in depressed patients. For example, anhedonia, the loss of pleasure, and abnormal motivation are key hallmarks of depression. In depression, the brain produces less dopamine and its receptors do not respond correctly, which underlies depressive symptoms such as anhedonia, demotivation, and changes in cognition. Antidepressants attempt to reverse these symptoms by targeting dopamine levels and receptors in the brain through various methods.

Some antidepressants reduce depressive symptoms by inhibiting dopamine reuptake (Papakostas, 2006). In other words, these drugs slow down the reabsorption of dopamine by dopamine transporters. This allows for dopamine to stay in the synapse for longer, increasing its effect on the brain. Since dopamine is associated with responding to pleasure and being motivated, which are not as present in depressed patients, slowing its reuptake can give rise to improvements in anhedonia and motivation issues in depression. One drug that inhibits dopamine reuptake is bupropion, which primarily blocks dopamine reuptake and also targets norepinephrine, another neurotransmitter in the brain (Ascher et al., 1995).

Other dopaminergic based antidepressants primarily inhibit the catabolism of dopamine –

also known as monoamine-oxidase inhibitors (MAOIs) (Papakostas, 2006). MAOIs target monoamine-oxidase (MAO), a family of enzymes that break down monoamines such as dopamine. This class of drugs blocks MAOs from typically functioning, thus prolonging dopamine's activity. Some examples of MAOIs include phenelzine, tranylcypromine and isocarboxazid.

Furthermore, some drugs increase dopamine activity in the brain by releasing dopamine and blocking its reuptake. These antidepressants are known as psychostimulants. Psychostimulants are used to treat depressed patients experiencing the symptoms of fatigue, hypersomnia, and apathy (Papakostas, 2006).

Serotonin Based Approaches

Serotonin is an important neurotransmitter in the brain. Serotonin plays a vital role in emotion, reproductive behavior, sleep, appetite, and many other processes (Kayabaşı et al., n.d.). The theory that deficits in serotonin give rise to depression has been around for over 50 years (Jauhar et al., 2023). Summarily, antidepressants that specifically target serotonin levels in the brain are a widely used approach for depression treatment.

A widely utilized serotonergic based class of drugs are selective serotonin reuptake inhibitors (SSRIs). SSRIs work by blocking the reuptake of serotonin into the cells, allowing the neurotransmitter to stay active in the brain for longer and therefore have a more pronounced effect on the brain (Dean & Keshavan, 2017). Since serotonin is important for regulating emotion and sleep patterns, SSRIs may be a helpful treatment for depressed patients experiencing varied emotions and sleep. SSRIs target only serotonin, not any other neurotransmitters like dopamine or norepinephrine, meaning that patients with abnormal serotonin levels would probably benefit the most from this medication. Also, SSRIs are effective at treating other mental disorders, such as anxiety (Gorman, 2000).

Furthermore, another group of effective serotonin based drugs is the serotonin / norepinephrine reuptake inhibitors (SNDRIs). Similar to SSRIs, SNDRIs work by stopping the reuptake of neurotransmitters by cells, allowing them to stay active and therefore having a greater effect on the brain. However, SNDRIs target the reuptake of both serotonin and norepinephrine (Gorman, 2000). Norepinephrine also helps regulate mood, and increasing norepinephrine via pharmacology has proven successful for depression treatment in some instances (Dean & Keshavan, 2017). Patients with irregular levels of norepinephrine and serotonin may experience more alleviation from depressive symptoms when taking SNDRIs.

Each antidepressant targets a specific mechanism that is functioning differently in a patient's brain. For example, if the dysfunction underlying an individual's depression is related to excessive serotonin reuptake, the individual would have the most successful results after taking a SSRI rather than a drug that targets dopamine reuptake. Gaining a better understanding of the differences in the neurochemistry underlying depression from patient to patient would make prescribing antidepressants and treatment overall more effective.

Talk Therapy

Talk therapy, also known as psychotherapy, is a variety of treatment methods used to help an individual overcome harmful emotions, thoughts, and behaviors (National Institute of Mental Health, n.d.). The goal of talk therapy is to alleviate patients' symptoms and equip the patient with coping mechanisms and logical framework for dealing with adverse instances which could affect them negatively. Talk therapy can treat a variety of mental illnesses, including depression. Therapy sessions are often directed by a licensed professional. These sessions can be held in either group settings or one-on-one meetings with a therapist. Psychotherapy has been proven to reduce depressive symptoms, demonstrating that therapy, in many cases, is an effective and primary treatment method for MDD (Das, 2024). Therapy can be used in tandem with or instead of antidepressants when treating depression (National Institute of Mental Health, n.d.).

A widely utilized version of talk therapy is cognitive behavioral therapy (CBT). Through CBT, patients learn how to manage their emotions by identifying and overcoming the thoughts and behaviors that underlie their depression (Reavell et al., 2018). For example, depressed individuals often experience a loss of confidence and overall negative outlook on life (Das, 2024). CBT may teach these individuals to deal with emotions more positively and effectively, relieving the individuals from these symptoms of depression.

Another type of psychotherapy is interpersonal therapy (IPT). IPT targets social dysfunction, which can be either a cause, effect, or sustaining factor of depression in an individual (Hirschfeld et al., 2000). IPT aims to treat depression by helping patients with their interpersonal relationships and skills (Papke et al., 2023). As stated before, depression is correlated with excessive amounts of stress. One such stressor may be social dysfunction in an individual's life. To alleviate themselves from this stress, the individual may attend IPT sessions to learn social skills and how to manage relationships better. Improved interpersonal relations can decrease the stress in an individual's life, which may decrease depressive symptoms.

However, psychotherapy has its limitations. Talk therapy is not completely effective all of the time and can be too weak of a method to treat a patient's depression (Strauss et al., 2021). In fact, some patients (5.2% of a sample of 14,587 patients) reported lasting negative effects of psychotherapy (Crawford et al., 2018). The benefits of therapy are not a guarantee, and this method does not work for everyone in all situations. Each individual requires a different type and amount of therapy to experience relief from depression, meaning that individualized care is necessary to improve treatment results.

Discussion

Major depressive disorder (MDD) is diagnosed by clinical professionals examining the symptom profile outlined in the DSM-5 TR. The factors underlying depression are extremely complicated and may involve the interaction of both genetic and environmental risk factors. Depression is associated with many structural changes in the brain, notably in the prefrontal cortex, amygdala, and hippocampus. Treatments for depression include antidepressants targeting dopamine and serotonin as well as talk therapy. Depression is a complex disorder, making

treatment equally complex. Due to this, many treatments are imperfect.

Because every patient has different underlying factors for their depression and varied anatomical structures in their brains, each patient requires a different assortment, amount, and type of treatment to truly see improvement. Also, the treatment process is based on trial and error, so it may take several months for an individual to find a method that is successful at lessening their depressive symptoms. This searching process may be grueling and disheartening, providing the opportunity for an individual's depression to become worse. Additionally, treatments may not be completely effective. Patients may still feel some symptoms of depression even while experiencing treatment. Studies demonstrate that only 20% of patients taking antidepressants see improvement in 6-8 weeks (Institute for Quality and Efficiency in Health Care [IQWiG], 2006).

To solve these issues and provide effective treatments for depression quickly, science needs to invest more into individualized care. The anatomical defects in each patient with depression are different, such as one person having hypoactivity in the prefrontal cortex while another has reduced volume in the hippocampus. If the same treatment is given to both individuals, it may not be effective in both scenarios since the underlying issue of each patient's depression is different. Each patient requires a treatment method unique to them that is correctly aimed at solving the issues in the patient's brain. Also, trial and error is not a sustainable method for treating depression, because there are a multitude of combinations and treatments available. For example, an individual suffering from hippocampal dysfunction and decreased volume can be prescribed drugs targeted toward the hippocampus along with a talk therapy approach designed to address emotional dysregulation and anhedonia. By combining talk therapy with pharmacology, it is a two-pronged approach converging on the same dysfunctional mechanism. This gives the patient a better opportunity to experience a more positive functional outcome.

An ideal treatment process would involve figuring out what the depressed individual is lacking, and then prescribing medicine and therapy sessions based on what the individual needs to restore their brain to its healthiest. It would also be beneficial to improve diagnostic criteria when providing treatment for depression. Observing the symptoms is an important first step, but understanding the circuitry issues underlying these symptoms is vital for providing accurate and effective depression treatment for patients. However, these two solutions are only possible if science gains a better understanding about the brain. Putting a larger emphasis on neuroscience research may be the key to adequately treating countless mental disorders including depression, which will improve the lives of many.

Works Cited

- Adolphs, R., Tranel, D., Damasio, H., & Damasio, A. R. (1995). Fear and the human amygdala. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 15(9), 5879–5891. <https://doi.org/10.1523/JNEUROSCI.15-09-05879.1995>
- Adolphs, R., Tranel, D., & Damasio, A. R. (1998). The human amygdala in social judgment. *Nature*, 393(6684), 470–474. <https://doi.org/10.1038/30982>
- Aggleton, J. P. (Ed.). (1992). *The amygdala: Neurobiological aspects of emotion, memory, and mental dysfunction*. Wiley-Liss.
- American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders* (5th ed., text rev.; DSM-5-TR). American Psychiatric Publishing. Ascher, J.A., Cole, J.O., Colin, J.N., Feighner, J.P., Ferris, R.M., Fibiger, H.C., Golden, R.N., Martin, P., Potter, W.Z., Richelson, E., 1995. Bupropion: a review of its mechanism of antidepressant activity. *J. Clin. Psychiatry* 56 (9), 395 – 401.
- Barth, J., Bermetz, L., Heim, E., Trelle, S., & Tonia, T. (2013). The current prevalence of child sexual abuse worldwide: a systematic review and meta-analysis. *International journal of public health*, 58(3), 469–483. <https://doi.org/10.1007/s00038-012-0426-1>
- Berton, O., & Nestler, E. J. (2006). New approaches to antidepressant drug discovery: beyond monoamines. *Nature reviews. Neuroscience*, 7(2), 137–151. <https://doi.org/10.1038/nrn1846>
- Bremner, J. D., Narayan, M., Anderson, E. R., Staib, L. H., Miller, H. L., & Charney, D. S. (2000). Hippocampal Volume Reduction in Major Depression. *American Journal of Psychiatry*, 157(1), 115–118. <https://doi.org/10.1176/ajp.157.1.115>
- Brown, G. W. (1989). Life events and measurement. In G. W. Brown & T. O. Harris (Eds.), *Life events and illness* (pp. 3–45). The Guilford Press.
- Burt, D. B., Zembar, M. J., & Niederehe, G. (1995). Depression and memory impairment: a meta-analysis of the association, its pattern, and specificity. *Psychological bulletin*, 117(2), 285–305. <https://doi.org/10.1037/0033-2909.117.2.285>
- Cardenas, D. D., Hoffman, J. M., Kirshblum, S., & McKinley, W. (2004). Etiology and incidence of rehospitalization after traumatic spinal cord injury: a multicenter analysis. *Archives of physical medicine and rehabilitation*, 85(11), 1757–1763. <https://doi.org/10.1016/j.apmr.2004.03.016>
- Chen, Z. Y., Patel, P. D., Sant, G., Meng, C. X., Teng, K. K., Hempstead, B. L., & Lee, F. S. (2004). Variant brain-derived neurotrophic factor (BDNF) (Met66) alters the intracellular trafficking and activity-dependent secretion of wild-type BDNF in neurosecretory cells and cortical neurons. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 24(18), 4401–4411. <https://doi.org/10.1523/JNEUROSCI.0348-04.2004>
- Dean, J., & Keshavan, M. (2017). The neurobiology of depression: An integrated view. *Asian Journal of Psychiatry*, 27, 101–111. <https://doi.org/10.1016/j.ajp.2017.01.025>
- Dougherty, D., & Rauch, S. L. (1997). Neuroimaging and neurobiological models of depression. *Harvard review of psychiatry*, 5(3), 138–159.

- <https://doi.org/10.3109/10673229709000299>
- Drevets, W. C., & Raichle, M. E. (1992). Neuroanatomical circuits in depression: implications for treatment mechanisms. *Psychopharmacology bulletin*, 28(3), 261–274.
- Dunn, R. T., Kimbrell, T. A., Ketter, T. A., Frye, M. A., Willis, M. W., Luckenbaugh, D. A., & Post, R. M. (2002). Principal components of the Beck Depression Inventory and regional cerebral metabolism in unipolar and bipolar depression. *Biological psychiatry*, 51(5), 387–399. [https://doi.org/10.1016/s0006-3223\(01\)01244-6](https://doi.org/10.1016/s0006-3223(01)01244-6)
- Ebmeier, K. P., Donaghey, C., & Steele, J. D. (2006). Recent developments and current controversies in depression. *Lancet (London, England)*, 367(9505), 153–167. [https://doi.org/10.1016/S0140-6736\(06\)67964-6](https://doi.org/10.1016/S0140-6736(06)67964-6)
- Etkin, A., Egner, T., Peraza, D. M., Kandel, E. R., & Hirsch, J. (2006). Resolving emotional conflict: a role for the rostral anterior cingulate cortex in modulating activity in the amygdala. *Neuron*, 51(6), 871–882. <https://doi.org/10.1016/j.neuron.2006.07.029>
- Gorman, K. (2000). SSRIs and SNRIs: Broad Spectrum of Efficacy Beyond Major Depression. *Physicians Postgraduate Press, Inc.*
- Hammen, C. (2018). Risk Factors for Depression: An Autobiographical Review. *Annual Review of Clinical Psychology*, 14(1), 1–28. <https://doi.org/10.1146/annurev-clinpsy-050817-084811>
- Hammen, C., Henry, R., & Daley, S. E. (2000). Depression and sensitization to stressors among young women as a function of childhood adversity. *Journal of Consulting and Clinical Psychology*, 68(5), 782–787. <https://doi.org/10.1037/0022-006X.68.5.782>
- Hathaway, W. R., & Newton, B. W. (2023). Neuroanatomy, Prefrontal Cortex. In StatPearls. StatPearls Publishing.
- Kayabaşı, Y., Güneş, B., & Erbaş, O. (2021). Serotonin Receptors and Depression. *Journal of Experimental and Basic Medical Sciences*.
- Kessler, R. C., Chiu, W. T., Demler, O., & Walters, E. E. (2005). Prevalence, Severity, and Comorbidity of 12-Month DSM-IV Disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, 62(6), 617. <https://doi.org/10.1001/archpsyc.62.6.617>
- Kolb, B., Mychasiuk, R., Muhammad, A., Li, Y., Frost, D. O., & Gibb, R. (2012). Experience and the developing prefrontal cortex. *Proceedings of the National Academy of Sciences*, 109(supplement_2), 17186–17193. <https://doi.org/10.1073/pnas.1121251109>
- Korgaonkar, M. S., Antees, C., Williams, L. M., Gatt, J. M., Bryant, R. A., Cohen, R., Paul, R., O'Hara, R., & Grieve, S. M. (2013). Early exposure to traumatic stressors impairs emotional brain circuitry. *PLoS ONE*, 8(9), e75524. <https://doi.org/10.1371/journal.pone.0075524>
- Lei, A. A., Phang, V. W. X., Lee, Y. Z., Kow, A. S. F., Tham, C. L., Ho, Y.-C., & Lee, M. T. (2025). Chronic Stress-Associated Depressive Disorders: The Impact of HPA Axis Dysregulation and Neuroinflammation on the Hippocampus—A Mini Review. *International Journal of Molecular Sciences*, 26(7), 2940. <https://doi.org/10.3390/ijms26072940>
- MacQueen, G., & Frodl, T. (2011). The hippocampus in major depression: evidence for the

- convergence of the bench and bedside in psychiatric research?. *Molecular psychiatry*, 16(3), 252–264. <https://doi.org/10.1038/mp.2010.80>
- Palazidou E. (2012). The neurobiology of depression. *British medical bulletin*, 101, 127–145. <https://doi.org/10.1093/bmb/lds004>
- Pandya, M., Altinay, M., Malone, D. A., & Anand, A. (2012). Where in the Brain Is Depression? *Current Psychiatry Reports*, 14(6), 634–642. <https://doi.org/10.1007/s11920-012-0322-7>
- Papakostas, G. I. (2006). Dopaminergic-based pharmacotherapies for depression. *European Neuropsychopharmacology*, 16(6), 391–402. <https://doi.org/10.1016/j.euroneuro.2005.12.002>
- Rigucci, S., Serafini, G., Pompili, M., Kotzalidis, G. D., & Tatarelli, R. (2010). Anatomical and functional correlates in major depressive disorder: the contribution of neuroimaging studies. *The world journal of biological psychiatry : the official journal of the World Federation of Societies of Biological Psychiatry*, 11(2 Pt 2), 165–180. <https://doi.org/10.1080/15622970903131571>
- Rolls, E. T. (2023). The prefrontal cortex. In E. T. Rolls, *Brain Computations and Connectivity* (2nd ed., pp. 596–623). Oxford University Press Oxford. <https://doi.org/10.1093/oso/9780198887911.003.0013>
- Saleh, A., Potter, G. G., McQuoid, D. R., Boyd, B., Turner, R., MacFall, J. R., & Taylor, W. D. (2017). Effects of early life stress on depression, cognitive performance and brain morphology. *Psychological medicine*, 47(1), 171–181. <https://doi.org/10.1017/S0033291716002403>
- Schmidt, H. D., Shelton, R. C., & Duman, R. S. (2011). Functional Biomarkers of Depression: Diagnosis, Treatment, and Pathophysiology. *Neuropsychopharmacology*, 36(12), 2375–2394. <https://doi.org/10.1038/npp.2011.151>
- Shadrina, M., Bondarenko, E. A., & Slominsky, P. A. (2018). Genetics Factors in Major Depression Disease. *Frontiers in Psychiatry*, 9, 334. <https://doi.org/10.3389/fpsy.2018.00334>
- Sheline, Y. I., Gado, M. H., & Price, J. L. (1998). Amygdala core nuclei volumes are decreased in recurrent major depression. *NeuroReport*, 9(9), 2023–2028. <https://doi.org/10.1097/00001756-199806220-00010>
- Siegle, G. J., Steinhauer, S. R., Thase, M. E., Stenger, V. A., & Carter, C. S. (2002). Can't shake that feeling: event-related fMRI assessment of sustained amygdala activity in response to emotional information in depressed individuals. *Biological psychiatry*, 51(9), 693–707. [https://doi.org/10.1016/s0006-3223\(02\)01314-8](https://doi.org/10.1016/s0006-3223(02)01314-8)
- Stahl, S. M. (n.d.). *Basic Psychopharmacology of Antidepressants, Part 1: Antidepressants Have Seven Distinct Mechanisms of Action*.
- Sullivan, P. F., Neale, M. C., & Kendler, K. S. (2000). Genetic epidemiology of major depression: review and meta-analysis. *The American journal of psychiatry*, 157(10), 1552–1562. <https://doi.org/10.1176/appi.ajp.157.10.1552>
- Swann, A. C., Steinberg, J. L., Lijffijt, M., & Moeller, F. G. (2008). Impulsivity: Differential

- relationship to depression and mania in bipolar disorder. *Journal of Affective Disorders*, 106(3), 241–248. <https://doi.org/10.1016/j.jad.2007.07.011>
- Thomas, K. M., Drevets, W. C., Dahl, R. E., Ryan, N. D., Birmaher, B., Eccard, C. H., Axelson, D., Whalen, P. J., & Casey, B. J. (2001). Amygdala response to fearful faces in anxious and depressed children. *Archives of general psychiatry*, 58(11), 1057–1063. <https://doi.org/10.1001/archpsyc.58.11.1057>
- Tolentino, J. C., & Schmidt, S. L. (2018). DSM-5 Criteria and Depression Severity: Implications for Clinical Practice. *Frontiers in Psychiatry*, 9. <https://doi.org/10.3389/fpsyt.2018.00450>
- Trifu, S. C., Trifu, A. C., Aluaș, E., Tătaru, M. A., & Costea, R. V. (2020). Brain changes in depression. *Romanian Journal of Morphology and Embryology*, 61(2), 361–370. <https://doi.org/10.47162/rjme.61.2.06>
- Volkow, N. D., Fowler, J. S., Wang, G. J., & Goldstein, R. Z. (2002). Role of dopamine, the frontal cortex and memory circuits in drug addiction: insight from imaging studies. *Neurobiology of learning and memory*, 78(3), 610–624. <https://doi.org/10.1006/nlme.2002.4099>
- Wang, Y.-B., Song, N.-N., Ding, Y.-Q., & Zhang, L. (2023). Neural plasticity and depression treatment. *IBRO Neuroscience Reports*, 14, 160–184. <https://doi.org/10.1016/j.ibneur.2022.09.001>
- Whalen, P. J., Kagan, J., Cook, R. G., Davis, F. C., Kim, H., Polis, S., McLaren, D. G., Somerville, L. H., McLean, A. A., Maxwell, J. S., & Johnstone, T. (2004). Human amygdala responsivity to masked fearful eye whites. *Science (New York, N.Y.)*, 306(5704), 2061. <https://doi.org/10.1126/science.1103617>
- Wright, J. H., & Beck, A. T. (1983). Cognitive Therapy of Depression: Theory and Practice. *Psychiatric Services*, 34(12), 1119–1127. <https://doi.org/10.1176/ps.34.12.1119>
- Zhao, F., Cheng, Z., Piao, J., Cui, R., & Li, B. (2022). Dopamine Receptors: Is It Possible to Become a Therapeutic Target for Depression? *Frontiers in Pharmacology*, 13. <https://doi.org/10.3389/fphar.2022.947785>

Rewiring Hope: The Promise of Adult Neurogenesis-Based Regenerative Therapies

By Avanish Srinivasan

Abstract

In the adult brain, neurons were long considered fixed and unchangeable. This central dogma of neurobiology was overturned in the late 20th century with the discovery of neurogenesis. This transformed the entire field and opened new therapeutic possibilities to the previously incurable neurodegenerative diseases (NDs). Evidence for neurogenesis has been found in various adult mammals, including humans, primarily in the subgranular zone (SGZ) and the subventricular zone (SVZ). In these regions, neural stem cells (NSCs) play a central role in regulating this process. NDs are vastly widespread. Alzheimer's disease alone affects 44 million people worldwide, while Parkinson's affects over 6 million, underscoring the desperate need for treatment options. NDs currently have no cure, with all treatments offering only meager symptomatic relief. Neurogenesis offers a novel and potentially transformative strategy, by restoring lost neurons and directly targeting the root cause of NDs. Neurogenesis-based therapy could treat and improve the outcome for millions, providing hope beyond symptom relief. This paper examines the biological foundations, current research, and future potential of neurogenesis as a viable therapeutic pathway. This is an urgent area of study given the growing aging population and burden of these diseases.

Introduction

For much of the 20th century, neuron generation was believed to have ended after embryonic development (Stahnisch). The possibility of new neuron formation was long considered unthinkable. Neuroscience was centered on the belief that neuron regeneration was impossible in adults. This idea stemmed primarily from the work of Santiago Ramón y Cajal (1852-1934). His experiments, utilizing Golgi staining (a method to visualize neuron morphology), observed the structure of neurons (Das et al.; Rozo et al.). He found that neurons are distinct, individual cells (Rozo et al.). This refuted the prevailing reticular theory, which held that the nervous system was a continuous network. Importantly, he observed no evidence of cell division, concluding that neurogenesis, or the formation of new neurons, was impossible. As he famously remarked, "In adult centers, the nerve paths are something fixed, ended, immutable. Everything may die, nothing may be regenerated." (Ramón y Cajal) This set the foundation for the core doctrine of neuroscience, ultimately forming the central dogma of neurobiology: that neurons can never regenerate (Colucci-D'Amato et al.).

This dogma seemed immutable throughout most of the 1900s (Stahnisch). Central Nervous System (CNS) injuries were thought to be irreversible, and adult neurogenesis was deemed impossible. This seemingly unshakable conviction was challenged in 1962 by Joseph Altman (Altman). Altman's seminal paper, *Are new neurons formed in the brains of adult mammals?*, presented evidence of neurogenesis in the brains of adult rats. When studying the proliferation, or division, of glia, or the support cells of neurons, Altman discovered the presence

of neurons and neuroblasts, cells that generate neurons, in the sites of trauma (Table 1) (Altman; Lin and Schagat). This implied that neurogenesis could indeed happen after a traumatic injury (Altman). Altman's discovery was widely controversial, and many disregarded it entirely, attributing it to errors in the experiment or faulty methodology (Porlan et al.).

After Altman's initial study, there was a surge of new studies related to neurogenesis. Notably, in the 1980s, Fernando Nottebohm demonstrated that neurogenesis in songbirds was linked to vocal learning and production (Table 1) (Nottebohm). A spike in neural cell count was seen in spring, when the songbirds sing and follow their courtship process. This proved that neurogenesis occurred in many species. This finding, much like Altman's, was largely dismissed, as many argued that evidence of neurogenesis in songbirds could not be applied to mammals or other species (Stahnisch). Such resistance reflected a broader reluctance to challenge the dogma. Yet, these findings across various species supported that neurogenesis was neither an isolated event nor an anomaly.

Over the following decades, a major technological innovation propelled the field forward: BrdU labeling (Table 1) (Martí-Clúa). Bromodeoxyuridine is a marker of DNA synthesis, which indicates mitosis, or cell division. BrdU is an excellent marker of adult neurogenesis, as it can identify and follow neuron progenitors (immature cells that can differentiate into mature neurons) in various species. BrdU labeling was used to identify neuron progenitors in adult rats and to provide evidence of adult neurogenesis in the human brain (Georg and Gage; Moreno-Jiménez et al.; Eriksson et al.). BrdU labeling became a pivotal technique in eventually overturning dogma.

The final catalyst for the shift in the paradigm was the 1990s' *Decade of the Brain*, which saw many prominent advancements in neuroscience research (Stahnisch). Arguably, the most important of these was the discovery of adult neural stem cells (NSCs). NSCs are undifferentiated cells that can differentiate into neurons and other brain cells (Zhao and Moore). Brent Reynolds and Samuel Weiss found the presence of multipotent NSCs in the brains of adult rats in their study in 1992 (Reynolds and Weiss). They proved that the mammalian brain could regenerate, starting from NSCs, and produce new neurons (Table 1). Together with the earlier findings by Altman and Nottebohm, these advancements confirmed the existence of adult neurogenesis and the significance of neuronal regeneration, ultimately helping dismantle the long-standing dogma.

With the help of new technology, Peter Eriksson and Fred Gage were able to finally provide irrefutable evidence of neurogenesis in 1998 (Eriksson et al.). Their findings confirmed the presence of neurogenesis in the dentate gyrus (DG) of the hippocampus in humans, definitively overturning the long-standing dogma of no new neurons (Table 1). Although previous studies in other species helped build the case, this study found BrdU-labeled neurons in humans, making it impossible to continue refuting adult neurogenesis. This marked a change in the paradigm after nearly a century. The historical trajectory of discoveries that led to this shift is outlined in Table 1.

Table 1: Timeline of advances in neuroscience and adult neurogenesis

Year	Milestone/Discovery	Author	Reference
1962	Adult mammalian neurogenesis in the injured rodent brain	Altman	(Altman)
1965	Neurogenesis identified in the DG of adult rats	Altman and Das	(Altman and Das)
1988	BrdU labeling was used to mark and trace cell proliferation	Miller and Nowakowski	(Miller and Nowakowski)
1989	Neurogenesis was identified in songbirds	Nottebohm	(Nottebohm)
1992	Discovered and isolated multipotent NSCs in the brains of adult rats	Reynolds and Weiss	(Reynolds and Weiss)
1997	Environmental factors were found to boost neurogenesis in the DG of mice	Kempermann et al.	(Kempermann et al.)
1998	Confirmed the presence of neurogenesis in the hippocampus of adult cancer patients	Eriksson et al.	(Eriksson et al.)
1999	Potential link between DG neurogenesis and learning was identified in adult rats	Gould et al.	(Gould et al.)
2004	Suggested presence of neurogenesis in the adult human OB	Bédard and Parent	(Bédard and Parent)
2007	Adult neurogenesis was observed in the human SVZ	Curtis et al.	(Curtis et al.)
2009	Loss of NSCs in adult mice led to impaired learning	Garthe et al.	(Garthe et al.)
2018	Strong evidence that neurogenesis persists with aging, with no sharp drop	Boldrini et al.	(Boldrini et al.)

This shift in belief opened new avenues for research in neuroscience, especially on the brain's innate ability to recover. With this understanding, researchers began to investigate the stimulation of neurogenesis as a potential therapeutic strategy to neurodegenerative diseases (NDs), offering renewed hope in a field once considered incurable (Howard). One of the key focuses of the therapeutic potential of neurogenesis lies in harnessing NSCs, the source of the newly generated neurons (Zhao and Moore). Stimulating differentiation could be key to controlling neurogenesis and its capacity to treat NDs.

Neurodegenerative disease is defined by neuron death (Lamprey et al.). The most common NDs include Alzheimer's disease (AD) and Parkinson's disease (PD). Other common NDs include Huntington's disease and Amyotrophic Lateral Sclerosis (ALS).

AD is the most prominent ND, characterized by the presence of amyloid- β (A β) (Bali et al.). A β is a fragment of a protein that forms plaques in the Alzheimer's-affected brain and

impairs cognitive function. This leads to the two most common symptoms: cognitive decline and memory loss. NSCs could be used as a potential treatment option, restoring the lost neurogenesis and neurons from AD (X. Chen et al.).

PD is another common ND, marked by impaired motor function, such as tremors, slow movement, and instability (Schulte and Gasser). The causes of PD remain widely debated. While some cases are linked to genetic factors, such as mutations to the LRRK2 gene, others can arise from environmental influences or from the accumulation of misfolded alpha-synuclein, a protein that becomes neurotoxic when misfolded (Schulte and Gasser; Müller-Nedebock et al.). A link between deregulated neurogenesis and the onset of PD has been seen, suggesting that the use of NSCs (to stimulate neurogenesis) could be a potential therapeutic strategy (Le Grand et al.).

While there are other neurodegenerative diseases, this paper will primarily focus on AD and PD, since they are the most prevalent NDs. AD affects approximately 44 million people worldwide, while PD affects over 6 million people (Caligiore et al.). NDs are currently incurable, and most therapeutic strategies offer only symptomatic relief (Palanisamy et al.). Stimulating neurogenesis, which could help restore neurons lost through NDs, presents a promising novel treatment strategy (Stahnisch). The recent advancements provide renewed hope and bear significant therapeutic implications. Therefore, this literature review will investigate how the discovery of adult neurogenesis and adult neural stem cells has advanced treatment strategies for neurodegenerative diseases.

Biology of adult neurogenesis and neural stem cells

Neurogenesis is defined as the generation of new neurons in the brain (Ming and Song). Historically, neurogenesis was believed to occur only during embryogenesis and the early postnatal development, with the adult brain lacking the capacity to generate new neurons. However, with Altman's groundbreaking studies, this idea was dispelled (Altman; Altman and Das). With the latest shift in the paradigm, research in the field has accelerated. The brain's potential for neurogenesis has been explored in many aspects, such as cellular pathways, regulatory mechanisms, and areas of activation.

Neuroanatomical foundations of neurogenesis

Clarifying the anatomical and functional basis of neurogenesis is critical to understanding how the discovery of NSCs can be harnessed as a therapeutic strategy for treating NDs. The brain is divided into two halves, called the left and right hemispheres (*Parts of the Brain*). Each hemisphere contains four sections, or lobes: frontal, parietal, temporal, and occipital (Figure 1). Inside the brain are four fluid-filled spaces known as ventricles, which help circulate cerebrospinal fluid. The surface of the brain is covered in folds, gyri (ridges), and sulci (grooves), which help increase the brain's surface area.

Within these structures lie specific areas of the brain responsible for neurogenesis. Neurogenesis is limited to these specific regions because these areas provide the necessary combination of NSCs, regulatory signals, and structural support needed to sustain the generation

of new neurons (Aimone et al.; Jurkowski et al.; Kempermann et al.; Y. Chen et al.). Some research suggests that the capacity is confined to regions where neurogenesis fulfills critical functions, such as memory, sensory input, and learning, and where neurons can be integrated without disrupting stability. Although there are other areas where neurogenesis may occur, active adult neurogenesis is primarily confined to two areas of the brain: the subgranular zone (SGZ) in the dentate gyrus (DG) of the hippocampus and the subventricular zone (SVZ) of the lateral ventricles (Figure 1) (Ming and Song).

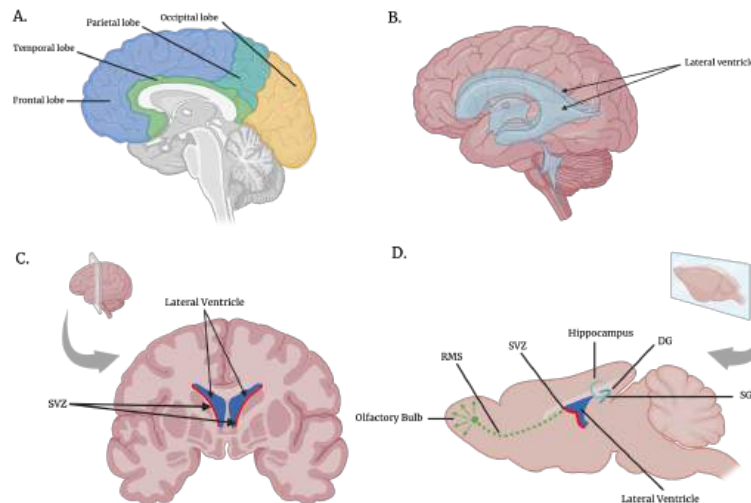


Figure 1: *Anatomical organization of major brain regions and limbic structures.* (A) Lateral view of the human cerebral cortex, highlighting the frontal lobe (blue), temporal lobe (green), parietal lobe (aqua), and occipital lobe (yellow). (B) Medial view of the human brain, showing the cingulate gyrus and corpus callosum in blue. (C) Coronal section of the human brain, illustrating the location of lateral ventricles in blue and the subventricular zone (SVZ) in red. (D) Medial temporal view of the mouse brain including the hippocampal formation, amygdala, and associated major input pathways.

Sites of active neurogenesis

The hippocampus, located in the temporal lobe, is a key site of neurogenesis (Ming and Song; *Parts of the Brain*). The hippocampus plays a vital role in learning and memory (*Parts of the Brain*). Within the hippocampus is the DG, aiding the process of memory formation (Amaral et al.). The SGZ is a specialized part of the DG that contains stem cells capable of differentiating into dentate granule cells, a type of neuron (Bressan and Saghatelian). The neurons produced migrate locally into the granule cell layer of the DG. This makes the SGZ a key site of adult neurogenesis (Ming and Song; Hsieh).

In addition to the SGZ, another major region where neurogenesis occurs is the SVZ (Conover and Allen). The SVZ lines the walls of the lateral ventricles. This region is the most prolific source of newly generated neurons in adulthood. The SVZ produces neuroblasts, which are immature cells that can become neurons (Martinez-Molina et al.). These neuroblasts migrate into other parts of the brain, where they differentiate into neurons.

Specifically, the SVZ is the primary source of neurons for the olfactory bulb (OB) (Martinez-Molina et al.; Kvistad et al.). Neuroblasts from the SVZ migrate a distance of approximately 3–8 millimeters toward the OB through the rostral migratory stream (RMS), a specialized migratory pathway between the SVZ and the forebrain (Martinez-Molina et al.). Once they reach the OB, the neuroblasts differentiate into interneurons (specialized neurons found between sensory and motor neurons) and disperse through the various layers of the OB (Martinez-Molina et al.; *Interneurons | Neuronetwork | University of Michigan Medical School*).

The OB is a vital structure in the brain responsible for maintaining olfactory (smell) information (Bayleyeg Derso et al.). Neurogenesis persists throughout an individual's lifetime in the OB, as it is vital for the functional plasticity of the olfactory neural circuits. The continual production of new neurons is important for maintaining olfactory function, including information processing and adaptation to environmental changes.

By comparison, hippocampal neurogenesis is strongly associated with cognition (Jurkowski et al.). In monkeys, a correlation between cognitive performance and levels of neurogenesis in the DG was observed. Similarly, human studies have shown a link between reduced neurogenic markers and age-related cognitive decline. These findings accentuate the importance of lifelong neurogenesis, not only for sensory systems but also for critical cognitive function. Together, these observations underscore the vital role of adult neurogenesis in maintaining neural plasticity and cognitive health.

While both the SGZ and SVZ are important sites of neurogenesis, research has historically favored the SVZ. There is a notable paucity of research on neurogenesis in the SGZ compared to the SVZ. This discrepancy may stem from many reasons, but the most plausible include the SVZ's anatomic accessibility, its robust neurogenic activity, and the dominance of rodent models (Alvarez-Buylla and García-Verdugo; Lois and Alvarez-Buylla; Cheng et al.; Mamber et al.; Zou et al.). As the largest site of neurogenesis, the SVZ is more accessible and has a well-defined migratory pathway. It also plays a role in olfaction, a critical sense for rodents, which may have contributed to a research bias due to the heavy reliance on rodent models.

Although neurogenesis in the mature brain occurs most prominently in the SGZ and SVZ, evidence suggests that some other regions of the brain may also possess a limited neurogenic capacity (Jurkowski et al.). However, these occurrences are often of a much smaller scale and remain a subject of ongoing scientific debate. Some of these novel neurogenic zones are the hypothalamus, amygdala, substantia nigra, striatum, cerebral cortex, and the third and fourth ventricles (Jurkowski et al.; Bayleyegn Derso et al.). These regions have shown signs of NSCs in some studies, suggesting that they may have a neurogenic capacity (Jurkowski et al.). However, they are not considered established neurogenic niches, or areas where neurogenesis occurs consistently, due to the lack of consistent evidence and activity. This observed neurogenic capacity in various regions raises the question about the characteristics of NSCs and the possibility that they could be found throughout the brain.

Properties and potential of neural stem cells

NSCs, or neural stem cells, have several distinct characteristics, and their behavior depends on the region where they are found (Ma et al.). In the brain, NSCs reside in a specialized microenvironment, known as the stem cell niche. The architecture of this niche determines the stem cell fate, such as whether it remains quiescent, differentiates, or proliferates. NSC behavior varies significantly depending on their niches. NSCs are multipotent, having the capacity to divide into multiple cell types, particularly neurons, astrocytes, and oligodendrocytes (Figure 2) (Zhao and Moore). Neurons, nerve cells that transmit information, are the primary targets of neurogenesis research (Zhao and Moore; *Brain Basics*). Neurons are the central focus because their regeneration holds the greatest therapeutic potential for restoring lost brain functions from diseases. The ability of NSCs to give rise to all three cell types underscores their function in maintaining neuroplasticity (Domingues et al.).

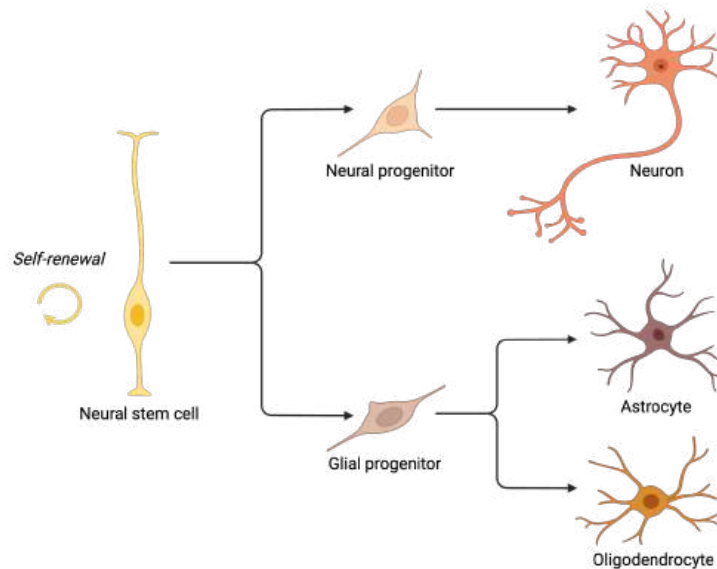


Figure 2: *Differentiation and self-renewal of neural stem cells into specialized neural cell types.* Neural stem cells (left) have the capacity for self-renewal and differentiation into neural progenitors. These progenitors have distinct developmental pathways: radial glial progenitors give rise to excitatory neurons (top pathway); glial progenitors produce astrocytes (middle pathway) which can help with neurotransmitter regulation. Glial progenitors also give rise to oligodendrocytes which form myelin sheaths around axons (bottom pathway). Arrows indicate lineage progression from multipotent stem cells to mature fate-committed cell types.

The morphology and behavior of the NSCs are contingent on their niche (Bayleyegn Derso et al.). Within the SGZ, NSCs are radial glial-like cells, often referred to as Type-1 cells. These cells are typically quiescent (dormant), dividing infrequently. When activated, they divide into intermediate progenitor cells, an immature cell with the potential to become a neuron. These progenitors exhibit doublecortin (DCX), a protein involved in neuronal migration, or polysialyated NCAM, a molecule essential for cell movement and migration (Bayleyegn Derso et al.; Gleeson et al.; Gómez-Climent et al.).

The intermediate progenitor cells have the potential to transition into neuroblasts, which

migrate through the hippocampal tissue, a process marked by the expression of cell adhesion molecules (Bayleyegn Derso et al.). These neuroblasts can then differentiate into dentate granule cells, mature cells of the DG, and become functionally integrated into existing neural circuits. In this final stage, post-mitotic neuronal markers, such as calcium-binding proteins, are expressed and used to identify this mature state. While this development of new neurons from NSCs applies to the SGZ, the SVZ has another example of neurogenesis in a different neural niche.

The SVZ is located on the lateral walls of the lateral ventricles and houses many NSCs (Figure 1). These NSCs, known as Type-B1 cells, resemble astrocytes in both morphology and molecular behavior. These cells express proteins such as GFAP, glutamate aspartate, and brain lipid-binding protein, commonly seen in astrocytes (Bayleyegn Derso et al.; Jurga et al.). Upon activation, Type-B1 cells begin expressing a different marker called nestin and divide asymmetrically (Ishiwata et al.; Bayleyegn Derso et al.). This means that one daughter cell retains stem cell identity, while another becomes a C cell, a type of progenitor cell (Gómez-López et al.; Bayleyegn Derso et al.). These C cells then proliferate, giving rise to neuroblasts that migrate along the RMS towards the OB, where they differentiate into mature interneurons (Bayleyegn Derso et al.). This sequence, NSC activation, proliferation, migration, and differentiation, mirrors the core stages of neurogenesis observed in the SGZ, highlighting the similarities across niches. Recognizing these patterns is crucial, as they may reflect shared regulatory mechanisms that could be utilized across niches for broader therapeutic applications.

Neural stem cells play a pivotal role in regulating the process of neurogenesis, indicating that modulation (targeted control) of NSCs could provide a means of controlling and harnessing neurogenesis (Zhao and Moore). One promising strategy to expand the NSC count *in vivo* (in living beings) and *in vitro* (outside living beings) involves the use of pluripotent stem cells, which can potentially be differentiated into NSCs. Human pluripotent stem cells (PSCs) are highly versatile cells capable of differentiating into nearly any cell type, making them invaluable in research (Romito and Cobellis). There are two types of PSCs: embryonic stem cells (ESCs), procured from embryos, and induced pluripotent stem cells (iPSCs), which are reprogrammed somatic (body) cells genetically modified to express stem cell properties (Zhao and Moore). Through neural induction, these PSCs can be activated and used to generate NSCs.

The ability to derive NSCs from pluripotent stem cells opens a novel avenue for developing therapeutic strategies. These PSC-derived NSCs provide hope for restoring lost cognitive function and repairing brain tissue damage from neurodegenerative diseases. More broadly, harnessing adult neurogenesis and NSCs could allow the restoration of abilities lost through aging, injury, and disease. Ultimately, expanding research and understanding of the biology of neurogenesis could help extend the frontiers of neuroscience knowledge and therapeutic capabilities.

Role of neurogenesis in neurodegenerative disorders

Neurogenesis is a fundamental process that plays a crucial role in maintaining cognitive function and plasticity. The impairment of neurogenesis has been increasingly associated with

the onset and progression of AD (Farioli-Vecchioli et al.). Particularly, a notable decline in adult hippocampal neurogenesis (AHN) has been observed. AHN is associated with memory and learning, and its diminution has been linked with the cognitive deficits of AD. Multiple studies, comprising both human postmortem and transgenic (altered genome with foreign genes) mice models, have seen AHN decline with aging and AD. Specifically, a decrease in NSC proliferation near A β plaques (protein plaquing seen in AD). This was also accompanied by elevated levels of BMP6, an inhibitor of neurogenesis (Farioli-Vecchioli et al.; Choi and Li). These findings suggest that disrupted neurogenesis, particularly in regions affected by AD, may be directly linked to cognitive decline characteristic of the disease. However, discrepancies exist amongst various studies regarding the presence and extent of neurogenesis impairment in AD, due to a lack of abundant research and methodology (Farioli-Vecchioli et al.). Overall, these findings demonstrate that AD actively suppresses neurogenesis and thereby aggravates the symptoms. This suggests that there is a link between the dysregulation of NSCs, impaired neurogenesis, and neurodegeneration.

Research suggests that PD may also involve comparable disturbances in neurogenesis and NSC function (Marxreiter et al.). Although PD has traditionally been associated with motor symptoms stemming from neuron loss in the substantia nigra, a growing array of non-motor symptoms has also been identified. These symptoms have been linked to the hippocampus and olfactory bulb. Both of these regions are key sites of adult neurogenesis. This supports a potential link between impaired neurogenesis and neurodegeneration from PD. Also, evidence from both postmortem analyses of human PD patients and rodent models has corroborated this notion. Particularly, studies found reduced dopaminergic innervation (dopamine-releasing nerve network) of the SVZ, indicated by a decrease in PCNA-positive cells, which act as markers of dopaminergic signaling (Marxreiter et al.; Parent et al.).

Dopamine has been shown to stimulate neurogenesis in the SVZ, and a reduction in dopaminergic innervation indicates impaired NSC proliferation. Additionally, a decrease in epidermal growth factor (EGF) levels was also seen (Marxreiter et al.; Boonstra et al.). EGF regulates cell proliferation, and a decrease suggests a reduction of OB neurogenesis. While these studies allude to a mechanistic link between PD and diminished neurogenesis, they remain somewhat controversial. This is because of substantial discrepancies across studies, which may stem from limitations such as the short preservation period of postmortem tissue, delays in tissue deterioration, antigen-masking, or methodology. Even so, preliminary evidence underscores the relevance of neurogenesis in PD pathology. Together, these findings accentuate the broader role of NSC dysregulation as a contributing factor to NDs.

The connection between NSC dysregulation and neurodegeneration is progressively being observed at the molecular level, particularly in epigenetic (how the environment can affect genetic expression) and post-transcriptional regulation (the stage between RNA transcription and translation) (Yen and Chen; “What Is Epigenetics?”; *Post-Transcriptional Control of Gene Expression* | *Biology for Majors I*). One such mechanism is N⁶-methyladenosine (m⁶A), the most prevalent and abundant mRNA modification in most eukaryotes (Yen and Chen). In NSCs, m⁶A

shapes gene expression necessary for proper neurodevelopment. The modifications are added through “writers”, such as METTL3, reversed through “erasers”, like FTO, and read by “readers”, for instance, the YHDF family. This epitranscriptome regulation, or the reversible modifications made to the RNA transcriptome, is vital in controlling NSC proliferation (Yen and Chen). Disruption of m⁶A leads to irregular behavior and impaired neurogenesis (Yen and Chen).

The dysregulation of NSCs is a hallmark of neurodegenerative conditions. In the brains affected by NDs, the m⁶A landscape is altered. “Readers” and “writers” are downregulated, changing the NSC proliferation rate and fate decisions. For example, the loss of METTL3 diminishes AHN and leads to many of the symptoms of early AD. Likewise, atypical expression of FTO is linked to many neurodegenerative conditions. Intrinsic stem cell dysfunction can help elucidate the neural decline observed in aging and NDs. Thus, harnessing m⁶A mechanisms might provide a novel therapeutic approach to NDs.

Utilizing NSC proliferation offers renewed therapeutic promise for the previously irremediable NDs. However, it is essential to dissect the regulatory networks that govern NSC fate to understand this strategy. NSCs are modulated by an array of regulatory mechanisms, each of which holds its importance. Among these, redox signaling and epigenetic modifications are particularly significant (Tay et al.). Redox (oxidation-reduction) reactions involve the transfer of electrons between two species (“Oxidation-Reduction Reactions”). Reactive oxygen species (ROS) are often the byproduct of redox reactions when oxygen is involved and act as a marker for redox (Tay et al.; Schieber and Chandel). Moderate increases in ROS act as a signal to activate quiescent NSCs, whereas excessive ROS damages NSC function (Tay et al.). Epigenetic modifiers, like histone-methyltransferase (HMT) and m⁶A, are influenced by the redox status. This redox-epigenetic crosstalk is a crucial regulator of NSC fate and behavior. Together, these findings highlight the importance of redox-epigenetic interactions in regulating the activity of NSCs, revealing a pathway that could be leveraged for neurogenesis-based therapies.

Transcriptional factors further contribute to NSC regulation (Guo et al.). For instance, Kruppel-like factor 9 (*Klf9*) functions as a negative regulator. The loss of *Klf9* leads to an expansion of the NSC population, suggesting that *Klf9* helps maintain NSC quiescence. Other transcriptional factors include *Erm*, *Sox9*, and *Ascl1* (Rajan et al.; Obernier and Alvarez-Buylla). On the extrinsic front, the Notch signaling pathway is a key mechanism that maintains NSC competency.

The mitochondria are another key component of NSC regulation (Coelho et al.). During differentiation, NSCs rely on a metabolic switch, from glycolysis to oxidative phosphorylation (OXPHOS). This switch is regulated by proteins such as PGC 1- α , where disruption leads to impaired neurogenesis and NSC proliferation. Collectively, these extrinsic and intrinsic mechanisms form a complex regulatory network that defines NSC biology. While many of the NSC regulatory mechanisms remain to be elucidated, the therapeutic potential of NSCs remains promising. Manipulation of these mechanisms will be essential for translating NSC biology into real-world therapeutic approaches for NDs.

Overview of ND treatment strategies

Contextualizing neurogenesis within the broader landscape of novel therapeutic strategies currently being researched is important for understanding its therapeutic potential. These emerging approaches seek to alter disease progression, not merely symptoms, like traditional methods. One such strategy is mesenchymal stem cell (MSC) therapy (Bhatt et al.). MSCs are a type of stromal cell (connective tissue cell) that can self-renew and differentiate into different cell types (Ding et al.; Manetti). They can cross the blood-brain barrier and exert therapeutic effects by secreting bioactive molecules, which promote neuroplasticity and angiogenesis (formation of new blood vessels) (Bhatt et al.; Adair and Montani). The therapy also reduces neuroinflammation, specifically helping clear A β plaques, offering a promising approach (Bhatt et al.). While MSCs support the existing neural environment, neurogenesis-based therapy directly targets the lost neurons from NDs, offering a more integrative approach.

Another promising therapeutic approach involves the use of antidiabetic agents, particularly GLP-1 receptor agonists and insulin sensitizers (Kopp et al.; Nowell et al.). These drugs were originally designed to help regulate glucose levels but can also be used to cross the blood-brain barrier and improve insulin signaling, which is often impaired in NDs. GLP-1 analogs have been shown to reduce A β plaque sizes and enhance cognitive function. Yet, compared to neurogenesis-based therapy, antidiabetics target the cellular environment rather than directly replenishing neurons, making them more supportive than curative. These strategies indicate a shift towards disease-modifying approaches, but compared to them, neurogenesis-based therapy uniquely aims to replace lost neurons, underscoring its distinct role in advancing treatment of NDs.

The ketogenic diet (KD) has also garnered attention as a therapeutic strategy, particularly for AD and PD (Dyńska et al.). KD promotes the production of ketone bodies (β -hydroxybutyrate, acetoacetate, and acetone), which serve as alternative energy sources for the brain. These metabolic changes may enhance neuronal survival and reduce neuroinflammation. Preliminary research has demonstrated KD's ability to improve cognitive function, motor symptoms, and disease progression, suggesting therapeutic potential. But, KD mainly offers metabolic support and symptomatic relief, rather than targeting the root cause of NDs: neuronal loss.

Finally, immunotherapy aims to modulate the immune system to reduce neuroinflammation and target proteins like A β (Mortada et al.). Vaccines could be used to induce an immune response that helps target ND proteins, showing therapeutic promise. However, despite encouraging preclinical studies, clinical trials have exhibited mixed outcomes. Compared to immunotherapies, which show variable efficacy, neurogenesis-based approaches or a combined approach offer greater potential for sustained recovery by restoring neuronal populations.

Neurogenesis-based strategies directly target the fundamental loss of neurons from NDs, offering a mechanism to restore rather than preserve. Unlike other novel approaches, neurogenesis holds the unique potential to help rebuild damaged circuits and recover cognitive function. This regenerative asset makes it arguably the most transformative avenue for therapy,

highlighting the need for further research to fully understand and harness its clinical applications.

Neurogenesis-based therapy

Given the limitations of current novel therapeutic strategies, neurogenesis-based therapy presents a fundamentally different and promising avenue. This approach targets the primary deficit from NDs: lost neural function and integration. By stimulating neurogenesis, cognitive function may be restored, and neuroregeneration could potentially be achieved.

Many different methods are being employed to treat neurodegenerative diseases using neurogenesis. The paper by Vassal et al. outlines the different methods in a comprehensive manner (Vassal et al.). But, briefly, several genetic approaches have shown promise. For instance, NSCs have been genetically modified to enhance their regenerative potential (Lutfi Ismaeel et al.). Another approach identified genes regulating NSC aging, using them to promote NSC longevity (Ruetz et al.). Lastly, restoring certain genes, such as miR-132, that have been reduced in NDs, helps improve neurogenesis and memory (Walgrave et al.).

Additionally, many stem cell transplantation techniques have demonstrated great potential (Shah et al.). NSC transplants have been shown to reduce neuroinflammation, enhance neuroplasticity, and improve cognition in AD models. Many clinical trials have used other cell types, such as MSCs, ESCs, and iPSCs, that are differentiated into NSCs to better replace lost neurons and support repair (Fan et al.). These transplanted cells may integrate into damaged regions, preserve existing neurons, or improve the brain's ability to heal (Chang et al.).

Numerous pharmacological therapies have also been used to promote neurogenesis. Compounds like IRL-1620 have been shown to enhance neuroregeneration and improve the Central Nervous System disorders (Gulati et al.). MFG-E8, a protein with many cellular functions, has been found to promote neurogenesis and protect neural tissue in disease models (Cheyuo et al.). Moreover, a neurosteroid, Allopregnanolone, has shown the ability to improve NSC proliferation and cognition in AD models (Brinton and Wang). Lastly, immunotherapy, or using the body's own immune system to fight disease, has also been explored to advance neurogenesis (Becker et al.; *What Is Immunotherapy?*). Anti-EFRH immunization in AD transgenic mouse models promoted neurogenesis and reduced disease pathology (Becker et al.).

Together, these strategies underscore the central role of neurogenesis and neural stem cells in advancing the treatment of NDs. By enhancing NSC function and proliferation, whether by genetic, transplantation, drug stimulation, or immune-based therapies, researchers aim to restore damaged circuits and replace lost neurons. These strategies move beyond traditional symptom-based therapy and offer potential to counter neuronal loss, the heart of NDs.

Discussion and future perspectives

Despite the significant advances in current understanding of neurogenesis, key gaps in knowledge hinder progress (Peretto and Bonfanti). Fundamental differences between embryonic and adult neurogenesis remain poorly understood (Hussain et al.). While the processes are clearly different, the mechanisms that lead to this divergence are unknown. Similarly, the

evolutionary disparities of neurogenesis remain unknown (Peretto and Bonfanti). Non-mammalian vertebrates use neurogenesis in a regenerative capacity, whereas in mammals, it serves a physiological purpose other than repair. The reason for this divergence remains speculative. The response to injury further accentuates this discrepancy: non-mammalian vertebrates exhibit a spike in neurogenesis and inflammation following injury, whilst mammals show impaired neurogenesis.

Moreover, the presence of neural progenitors is also not a guarantee of neurogenesis. Only specific progenitor cells possess neurogenic capacity. Other unresolved questions concern non-traditional sites of neurogenesis. The existence, regulation, and functional pertinence of neurogenesis in these regions remain unclear. Additionally, while other forms of neuroplasticity occur in the brain, the relationship between neurogenesis and these mechanisms is largely unexplored.

A major challenge in the field is translating findings from animal models, particularly rodents, to humans. Most research on adult neurogenesis has been centered on rodents, whose brains, although strikingly similar, differ from humans in both structure and function. For instance, while RMS migration persists throughout life in rodents, it is markedly limited in the postnatal human brain. Rodents also show a steady age-related decline in the rate of neurogenesis, whereas humans exhibit a more consistent rate. Differences in lifespan, maturation, and developmental timing further complicate translational efforts. Finally, many experimental methods used in rodents are not feasible in humans due to technical and ethical constraints, leading to a limited clinical applicability of findings (Peretto and Bonfanti; Jessberger and Gage).

These unresolved questions, along with species-to-species differences and methodological limitations, continue to present challenges for the field. Addressing them is essential for improving the translational relevance of current findings and for advancing neurogenesis-based therapeutic approaches.

Risks and ethical challenges in neurogenesis research & future outlooks

A fundamental component of neurogenesis-based therapies is the use of stem cells. However, the transplantation of NSCs or other neural progenitors carries significant risks for the recipient (Piscaglia). One of the most serious concerns is tumorigenesis, or the potential for the transplanted cells to form tumors. Stem cells can acquire mutations that lead to uncontrolled tumor formation. As a result, safety remains a barrier to clinical translation.

Another major challenge is immune rejection, where the recipient's immune system may attack the transplanted cells (Piscaglia; Herberts et al.). This outcome can vary widely between patients. A further risk is unwanted differentiation, in which the stem cells differentiate into unintended cell types (Herberts et al.). For instance, MSCs have been observed to differentiate into bone or cartilage, which could have severe and harmful effects depending on the site of transplantation. Additionally, the relative risks and therapeutic benefits of the different stem cell types are not well understood, highlighting the need for further research and evaluation.

(Hachimi-Idrissi).

Stem cell research and its clinical application entail a range of ethical considerations, particularly concerning the use of embryos and ESCs (Lo and Parham). A central controversy centers on the moral acceptability of using human embryos for research purposes. Many debates are often focused on the definition and onset of personhood. Additional ethical issues arise during clinical trials, where participants misinterpret the objectives of the early-phase therapies, frequently overestimating their efficacy and misconstruing the studies as therapeutic rather than safety evaluations.

Neurogenesis-based strategies introduce further ethical challenges (Alex et al.). Potential side effects of therapy, such as genetic mutations and unintended alterations to behavior, raise ethical concerns about the long-term consequences of such therapies. Moreover, the use of scanning and labeling techniques, such as MRI and BrdU administration, in human participants prompts further inspection over the ethical boundaries of human experimentation (Alex et al.; Lieberwirth and Wang). Moving forward, it is of paramount importance that research protocols emphasize transparency, consent, and long-term monitoring of participants to preserve ethical integrity (Lo and Parham; Alex et al.). Addressing these ethical challenges will be essential to responsibly advancing the therapeutic potential of neurogenesis.

The discovery of neurogenesis has reshaped the idea of neuroplasticity, providing insight into the brain's complex mechanism of recovery. As neurodegenerative diseases continue to affect millions, neurogenesis and neural stem cells provide an exciting direction for therapeutic practices in the future (Kim et al.). Overall, neurogenesis provides hope to restore cognitive impairments, support learning, and improve memory formation in patients.

Nevertheless, this optimism must be matched with pragmatism. Although neurogenesis may help recover aspects of cognitive function, it does not entail the recovery of memories that have already been lost (Lazarov et al.). This is because the neurons encoding these memories may have already died as a result of disease, underscoring a major limitation to therapeutic approaches utilizing neurogenesis. Some research suggests NDs, like AD, impair the ability to recall memories rather than destroying neurons completely (Perusini et al.). However, even in such cases, neurogenesis-based therapies cannot currently address these underlying deficits in memory recall. Although neurogenesis cannot restore memories that have already been lost, it can still enhance aspects of daily functioning in patients, such as memory formation, learning, and mood regulation. These improvements, while valuable, are accompanied by a lingering absence of personal memories that form the core identity of an individual. Consequently, neurogenesis-based therapies improve the quality of life but fall short of fully restoring one's livelihood.

Significant challenges also remain. Both physiological and ethical constraints hinder further progress in research. For instance, risks associated with stem cell transplants and the ethical implications of manipulating human stem cells continue to limit clinical applications. However, these challenges can be addressed through transparency and the development of rigorous ethical guidelines that can guide future research in the field. This is especially important

as neuroscience currently lacks a well-established ethical framework, which is vital for progress (Sawai and Kataoka). Despite these caveats, the recent breakthroughs in the field open new avenues in therapeutics for the previously incurable NDs. While obstacles remain, the persistence of investigation into the issue reflects a broader hope: that neurogenesis may one day help cure the cognitive deficits of neurodegenerative disease.

Looking ahead, several novel avenues merit exploration. One potential study could utilize AI-driven brain activity mapping. This could be used to identify patient-specific regions of degeneration and determine the optimal sites of stem cell implantation. This approach could not only enhance precision but also provide a method to improve integration. Another unexplored concept is the idea of cognitive priming before or after NSC therapy. Cognitive priming is engaging in cognitive tasks, such as classifying items in relation to previous items, which can help stimulate learning and memory formation (Schacter et al.). Such priming may help facilitate the integration of new neurons or boost the overall efficacy of treatment. These ideas represent a fraction of the unexplored intersections between different ideas in neurobiology. As research advances, incorporating biological insight with technological innovations may hold the key to treating neurodegenerative disease.

Works Cited

- Adair, Thomas H., and Jean-Pierre Montani. "Overview of Angiogenesis." *Angiogenesis*, Morgan & Claypool Life Sciences, 2010. www.ncbi.nlm.nih.gov, <https://www.ncbi.nlm.nih.gov/books/NBK53238/>.
- Aimone, James B., et al. "Regulation and Function of Adult Neurogenesis: From Genes to Cognition." *Physiological Reviews*, vol. 94, no. 4, Oct. 2014. Europe PMC, <https://doi.org/10.1152/physrev.00004.2014>.
- Alex, Gideon S., et al. "Adult Neurogenesis: A Review of Current Perspectives and Implications for Neuroscience Research." *Journal of Neuroscience and Neurological Disorders*, vol. 8, no. 2, Nov. 2024, pp. 106–14. www.neuroscijournal.com, <https://doi.org/10.29328/journal.jnnd.1001102>.
- Altman, Joseph. "Are New Neurons Formed in the Brains of Adult Mammals?" *Science*, vol. 135, no. 3509, Mar. 1962, pp. 1127–28. science.org (Atypon), <https://doi.org/10.1126/science.135.3509.1127>.
- Altman, Joseph, and Gopal D. Das. "Autoradiographic and Histological Evidence of Postnatal Hippocampal Neurogenesis in Rats." *Journal of Comparative Neurology*, vol. 124, no. 3, 1965, pp. 319–35. Wiley Online Library, <https://doi.org/10.1002/cne.901240303>.
- Alvarez-Buylla, Arturo, and Jose Manuel García-Verdugo. "Neurogenesis in Adult Subventricular Zone." *The Journal of Neuroscience*, vol. 22, no. 3, Feb. 2002, pp. 629–34. PubMed Central, <https://doi.org/10.1523/JNEUROSCI.22-03-00629.2002>.
- Amaral, David G., et al. "The Dentate Gyrus: Fundamental Neuroanatomical Organization (Dentate Gyrus for Dummies)." *Progress in Brain Research*, vol. 163, 2007, pp. 3–22. PubMed Central, [https://doi.org/10.1016/S0079-6123\(07\)63001-5](https://doi.org/10.1016/S0079-6123(07)63001-5).
- Bali, Jitin, et al. "Cellular Basis of Alzheimer's Disease." *Annals of Indian Academy of Neurology*, vol. 13, no. Suppl2, Dec. 2010, pp. S89–93. PubMed Central, <https://doi.org/10.4103/0972-2327.74251>.
- Bayleyegn Derso, Teketay, et al. "Neural Stem Cells in Adult Neurogenesis and Their Therapeutic Applications in Neurodegenerative Disorders: A Concise Review." *Frontiers in Molecular Medicine*, vol. 5, June 2025. Frontiers, <https://doi.org/10.3389/fmmed.2025.1569717>.
- Becker, Maria, et al. "Stimulation of Endogenous Neurogenesis by Anti-EFRH Immunization in a Transgenic Mouse Model of Alzheimer's Disease." *Proceedings of the National Academy of Sciences*, vol. 104, no. 5, Jan. 2007, pp. 1691–96. pnas.org (Atypon), <https://doi.org/10.1073/pnas.0610180104>.
- Bhatt, Aditya, et al. "Mesenchymal Stem Cell Therapy for Alzheimer's Disease: A Novel Therapeutic Approach for Neurodegenerative Diseases." *Neuroscience*, vol. 555, Sept. 2024, pp. 52–68. www.ibroneuroscience.org, <https://doi.org/10.1016/j.neuroscience.2024.07.019>.
- Boonstra, J., et al. "The Epidermal Growth Factor." *Cell Biology International*, vol. 19, no. 5, May 1995, pp. 413–30. PubMed, <https://doi.org/10.1006/cbir.1995.1086>.
- Brain Basics: The Life and Death of a Neuron | National Institute of Neurological Disorders and Stroke. <https://www.ninds.nih.gov/health-information/public-education/brain-basics/brain-basics->

- life-and-death-neuron. Accessed 18 July 2025.
- Bressan, Cedric, and Armen Saghatelian. "Intrinsic Mechanisms Regulating Neuronal Migration in the Postnatal Brain." *Frontiers in Cellular Neuroscience*, vol. 14, Jan. 2021. Frontiers, <https://doi.org/10.3389/fncel.2020.620379>.
- Brinton, Roberta Diaz, and Jun Ming Wang. "Therapeutic Potential of Neurogenesis for Prevention and Recovery from Alzheimer's Disease: Allopregnanolone as a Proof of Concept Neurogenic Agent." *Current Alzheimer Research*, vol. 3, no. 3, July 2006, pp. 185–90. PubMed, <https://doi.org/10.2174/156720506777632817>.
- Caligiore, Daniele, et al. "The Neurodegenerative Elderly Syndrome (NES) Hypothesis: Alzheimer and Parkinson Are Two Faces of the Same Disease." *IBRO Neuroscience Reports*, vol. 13, Dec. 2022, pp. 330–43. ScienceDirect, <https://doi.org/10.1016/j.ibneur.2022.09.007>.
- Chang, Jun, et al. "Neural Stem Cells Promote Neuroplasticity: A Promising Therapeutic Strategy for the Treatment of Alzheimer's Disease." *Neural Regeneration Research*, vol. 19, no. 3, July 2023, pp. 619–28. PubMed Central, <https://doi.org/10.4103/1673-5374.380874>.
- Chen, Xiaokun, et al. "Neural Stem Cells in the Treatment of Alzheimer's Disease: Current Status, Challenges, and Future Prospects." *Journal of Alzheimer's Disease*, vol. 94, no. Suppl 1, pp. S173–86. PubMed Central, <https://doi.org/10.3233/JAD-220721>.
- Chen, Yali, et al. "Olfactory Bulb Neurogenesis Depending on Signaling in the Subventricular Zone." *Cerebral Cortex (New York, N.Y.: 1991)*, vol. 33, no. 22, Nov. 2023, pp. 11102–11. PubMed, <https://doi.org/10.1093/cercor/bhad349>.
- Cheng, Li-Chun, et al. "miR-124 Regulates Adult Neurogenesis in the Subventricular Zone Stem Cell Niche." *Nature Neuroscience*, vol. 12, no. 4, Apr. 2009, pp. 399–408. PubMed, <https://doi.org/10.1038/nn.2294>.
- Cheyuo, Cletus, et al. "Neurogenesis in Neurodegenerative Diseases: Role of MFG-E8." *Frontiers in Neuroscience*, vol. 13, June 2019. Frontiers, <https://doi.org/10.3389/fnins.2019.00569>.
- Choi, Se Hoon, and Yun Li. "Elevated Levels of BMP6 Impair Neurogenesis in Alzheimer's Disease." *The Journal of Neuroscience*, vol. 31, no. 2, Jan. 2011, pp. 371–72. PubMed Central, <https://doi.org/10.1523/JNEUROSCI.5226-10.2011>.
- Coelho, Patrícia, et al. "Mitochondrial Function and Dynamics in Neural Stem Cells and Neurogenesis: Implications for Neurodegenerative Diseases." *Ageing Research Reviews*, vol. 80, Sept. 2022, p. 101667. ScienceDirect, <https://doi.org/10.1016/j.arr.2022.101667>.
- Colucci-D'Amato, L., et al. "The End of the Central Dogma of Neurobiology: Stem Cells and Neurogenesis in Adult CNS." *Neurological Sciences: Official Journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology*, vol. 27, no. 4, Sept. 2006, pp. 266–70. PubMed, <https://doi.org/10.1007/s10072-006-0682-z>.
- Conover, J. C., and R. L. Allen. "The Subventricular Zone: New Molecular and Cellular Developments." *Cellular and Molecular Life Sciences: CMLS*, vol. 59, no. 12, Dec. 2002, pp. 2128–35. PubMed Central, <https://doi.org/10.1007/s000180200012>.
- Das, Gitanjali, et al. "The Golgi-Cox Method." *Methods in Molecular Biology (Clifton, N.J.)*,

- vol. 1018, 2013, pp. 313–21. PubMed, https://doi.org/10.1007/978-1-62703-444-9_29.
- Ding, Dah-Ching, et al. “Mesenchymal Stem Cells.” *Cell Transplantation*, vol. 20, no. 1, 2011, pp. 5–14. PubMed, <https://doi.org/10.3727/096368910X>.
- Domingues, Helena S., et al. “Oligodendrocyte, Astrocyte, and Microglia Crosstalk in Myelin Development, Damage, and Repair.” *Frontiers in Cell and Developmental Biology*, vol. 4, June 2016, p. 71. PubMed Central, <https://doi.org/10.3389/fcell.2016.00071>.
- Dyńska, Damian, et al. “The Role of Ketogenic Diet in the Treatment of Neurological Diseases.” *Nutrients*, vol. 14, no. 23, Nov. 2022, p. 5003. PubMed Central, <https://doi.org/10.3390/nu14235003>.
- Epitranscriptome - an Overview | ScienceDirect Topics.
<https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/epitranscriptome>. Accessed 24 July 2025.
- Eriksson, Peter S., et al. “Neurogenesis in the Adult Human Hippocampus.” *Nature Medicine*, vol. 4, no. 11, Nov. 1998, pp. 1313–17. www.nature.com, <https://doi.org/10.1038/3305>.
- Fan, Yiping, et al. “Neural Cells for Neurodegenerative Diseases in Clinical Trials.” *Stem Cells Translational Medicine*, vol. 12, no. 8, July 2023, pp. 510–26. PubMed Central, <https://doi.org/10.1093/stcltm/szad041>.
- Farioli-Vecchioli, Stefano, et al. “Adult Hippocampal Neurogenesis in Alzheimer’s Disease: An Overview of Human and Animal Studies with Implications for Therapeutic Perspectives Aimed at Memory Recovery.” *Neural Plasticity*, vol. 2022, Jan. 2022, p. 9959044. PubMed Central, <https://doi.org/10.1155/2022/9959044>.
- Georg, H, and Fred H Gage. Neurogenesis in the Dentate Gyrus of the Adult Rat: Age-Related Decrease of Neuronal Progenitor Proliferation.
- Gleeson, J. G., et al. “Doublecortin Is a Microtubule-Associated Protein and Is Expressed Widely by Migrating Neurons.” *Neuron*, vol. 23, no. 2, June 1999, pp. 257–71. PubMed, [https://doi.org/10.1016/s0896-6273\(00\)80778-3](https://doi.org/10.1016/s0896-6273(00)80778-3).
- Gómez-Climent, María Ángeles, et al. “The Polysialylated Form of the Neural Cell Adhesion Molecule (PSA-NCAM) Is Expressed in a Subpopulation of Mature Cortical Interneurons Characterized by Reduced Structural Features and Connectivity.” *Cerebral Cortex* (New York, N.Y.: 1991), vol. 21, no. 5, May 2011, pp. 1028–41. PubMed, <https://doi.org/10.1093/cercor/bhq177>.
- Gómez-López, Sandra, et al. “Asymmetric Cell Division of Stem and Progenitor Cells during Homeostasis and Cancer.” *Cellular and Molecular Life Sciences: CMLS*, vol. 71, no. 4, June 2013, pp. 575–97. PubMed Central, <https://doi.org/10.1007/s00018-013-1386-1>.
- Gulati, A., et al. “A Novel Neuroregenerative Approach Using ETB Receptor Agonist, IRL-1620, to Treat CNS Disorders.” *Physiological Research*, May 2018, pp. S95–113. DOI.org (Crossref), <https://doi.org/10.33549/physiolres.933859>.
- Guo, Nannan, et al. “Transcriptional Regulation of Neural Stem Cell Expansion in the Adult Hippocampus.” *eLife*, vol. 11, p. e72195. PubMed Central, <https://doi.org/10.7554/eLife.72195>.

- Hachimi-Idrissi, Said. “Stem Cell Therapy in Neurological Disorders: Promises and Concerns.” *Exploration of Neuroprotective Therapy*, vol. 3, no. 5, no. 5, Oct. 2023, pp. 346–62. www.explorationpub.com, <https://doi.org/10.37349/ent.2023.00055>.
- Herberts, Carla A., et al. “Risk Factors in the Development of Stem Cell Therapy.” *Journal of Translational Medicine*, vol. 9, no. 1, Mar. 2011, p. 29. BioMed Central, <https://doi.org/10.1186/1479-5876-9-29>.
- Howard, Ken. Princeton - PWB 040599 - Do Brain Cells Regenerate? 5 Apr. 1999, <https://pr.princeton.edu/pwb/99/0405/brain.htm>.
- Hsieh, Jenny. “Orchestrating Transcriptional Control of Adult Neurogenesis.” *Genes & Development*, vol. 26, no. 10, May 2012, pp. 1010–21. PubMed Central, <https://doi.org/10.1101/gad.187336.112>.
- Hussain, Ghulam, et al. “Adult Neurogenesis: A Real Hope or a Delusion?” *Neural Regeneration Research*, vol. 19, no. 1, May 2023, pp. 6–15. PubMed Central, <https://doi.org/10.4103/1673-5374.375317>.
- Interneurons | Neuronetwork | University of Michigan Medical School. <https://medresearch.umich.edu/labs-departments/centers/mneuronet/research/interneurons>. Accessed 17 July 2025.
- Ishiwata, Toshiyuki, et al. “Neuroepithelial Stem Cell Marker Nestin Regulates the Migration, Invasion and Growth of Human Gliomas.” *Oncology Reports*, vol. 26, no. 1, July 2011, pp. 91–99. PubMed, <https://doi.org/10.3892/or.2011.1267>.
- Jessberger, Sebastian, and Fred H. Gage. “Adult Neurogenesis: Bridging the Gap between Mice and Humans.” *Trends in Cell Biology*, vol. 24, no. 10, Oct. 2014, pp. 558–63. www.cell.com, <https://doi.org/10.1016/j.tcb.2014.07.003>.
- Jurga, Agnieszka M., et al. “Beyond the GFAP-Astrocyte Protein Markers in the Brain.” *Biomolecules*, vol. 11, no. 9, Sept. 2021, p. 1361. PubMed Central, <https://doi.org/10.3390/biom11091361>.
- Jurkowski, Michal P., et al. “Beyond the Hippocampus and the SVZ: Adult Neurogenesis Throughout the Brain.” *Frontiers in Cellular Neuroscience*, vol. 14, Sept. 2020. Frontiers, <https://doi.org/10.3389/fncel.2020.576444>.
- Kempermann, Gerd, et al. “Neurogenesis in the Adult Hippocampus.” *Cold Spring Harbor Perspectives in Biology*, vol. 7, no. 9, Sept. 2015, p. a018812. PubMed Central, <https://doi.org/10.1101/cshperspect.a018812>.
- Kim, Thomas A., et al. “Adult Hippocampal Neurogenesis and Its Impairment in Alzheimer’s Disease.” *Zoological Research*, vol. 43, no. 3, May 2022, pp. 481–96. PubMed Central, <https://doi.org/10.24272/j.issn.2095-8137.2021.479>.
- Kopp, Katherine O., et al. “Glucagon-like Peptide-1 (GLP-1) Receptor Agonists and Neuroinflammation: Implications for Neurodegenerative Disease Treatment.” *Pharmacological Research*, vol. 186, Dec. 2022, p. 106550. PubMed, <https://doi.org/10.1016/j.phrs.2022.106550>.
- Kvistad, Christopher Elnan, et al. “Neural Regeneration in the Human Central Nervous

- System—from Understanding the Underlying Mechanisms to Developing Treatments. Where Do We Stand Today?” *Frontiers in Neurology*, vol. 15, May 2024. *Frontiers*, <https://doi.org/10.3389/fneur.2024.1398089>.
- Lampitey, Richard N. L., et al. “A Review of the Common Neurodegenerative Disorders: Current Therapeutic Approaches and the Potential Role of Nanotherapeutics.” *International Journal of Molecular Sciences*, vol. 23, no. 3, Feb. 2022, p. 1851. *PubMed Central*, <https://doi.org/10.3390/ijms23031851>.
- Lazarov, Orly, et al. “Memory Circuits in Dementia: The Engram, Hippocampal Neurogenesis and Alzheimer’s Disease.” *Progress in Neurobiology*, vol. 236, May 2024, p. 102601. *ScienceDirect*, <https://doi.org/10.1016/j.pneurobio.2024.102601>.
- Le Grand, Jaclyn Nicole, et al. “Neural Stem Cells in Parkinson’s Disease: A Role for Neurogenesis Defects in Onset and Progression.” *Cellular and Molecular Life Sciences: CMLS*, vol. 72, no. 4, Nov. 2014, pp. 773–97. *PubMed Central*, <https://doi.org/10.1007/s00018-014-1774-1>.
- Lieberwirth, Claudia, and Zuoxin Wang. “The Social Environment and Neurogenesis in the Adult Mammalian Brain.” *Frontiers in Human Neuroscience*, vol. 6, May 2012, p. 118. *PubMed Central*, <https://doi.org/10.3389/fnhum.2012.00118>.
- Lin, H., and T. Schagat. “Neuroblasts: A Model for the Asymmetric Division of Stem Cells.” *Trends in Genetics: TIG*, vol. 13, no. 1, Jan. 1997, pp. 33–39. *PubMed*, [https://doi.org/10.1016/s0168-9525\(96\)10050-0](https://doi.org/10.1016/s0168-9525(96)10050-0).
- Lo, Bernard, and Lindsay Parham. “Ethical Issues in Stem Cell Research.” *Endocrine Reviews*, vol. 30, no. 3, May 2009, pp. 204–13. *PubMed Central*, <https://doi.org/10.1210/er.2008-0031>.
- Lois, C., and A. Alvarez-Buylla. “Long-Distance Neuronal Migration in the Adult Mammalian Brain.” *Science (New York, N.Y.)*, vol. 264, no. 5162, May 1994, pp. 1145–48. *PubMed*, <https://doi.org/10.1126/science.8178174>.
- Lutfi Ismaeel, Ghufraan, et al. “Genetically Engineered Neural Stem Cells (NSCs) Therapy for Neurological Diseases; State-of-the-Art.” *Biotechnology Progress*, vol. 39, no. 5, 2023, p. e3363. *PubMed*, <https://doi.org/10.1002/btpr.3363>.
- Ma, Dengke K., et al. “Adult Neural Stem Cells in the Mammalian Central Nervous System.” *Cell Research*, vol. 19, no. 6, June 2009, pp. 672–82. *PubMed Central*, <https://doi.org/10.1038/cr.2009.56>.
- Mamber, Carlyn, et al. “Shades of Gray: The Delineation of Marker Expression within the Adult Rodent Subventricular Zone.” *Progress in Neurobiology*, vol. 111, Dec. 2013, pp. 1–16. *ScienceDirect*, <https://doi.org/10.1016/j.pneurobio.2013.07.003>.
- Manetti, Mirko. “Molecular Morphology and Function of Stromal Cells.” *International Journal of Molecular Sciences*, vol. 22, no. 24, Dec. 2021, p. 13422. *PubMed Central*, <https://doi.org/10.3390/ijms222413422>.
- Martí-Clúa, Joaquín. “5-Bromo-2’-Deoxyuridine Labeling: Historical Perspectives, Factors Influencing the Detection, Toxicity, and Its Implications in the Neurogenesis.” *Neural*

- Regeneration Research, vol. 19, no. 2, July 2023, p. 302. [pmc.ncbi.nlm.nih.gov, https://doi.org/10.4103/1673-5374.379038](https://doi.org/10.4103/1673-5374.379038).
- Martinez-Molina, Noelia, et al. “Rostral Migratory Stream Neuroblasts Turn and Change Directions in Stereotypic Patterns.” *Cell Adhesion & Migration*, vol. 5, no. 1, 2011, pp. 83–95. PubMed Central, <https://doi.org/10.4161/cam.5.1.13788>.
- Marxreiter, Franz, et al. “Adult Neurogenesis in Parkinson’s Disease.” *Cellular and Molecular Life Sciences: CMLS*, vol. 70, no. 3, July 2012, pp. 459–73. PubMed Central, <https://doi.org/10.1007/s00018-012-1062-x>.
- Ming, Guo-li, and Hongjun Song. “Adult Neurogenesis in the Mammalian Brain: Significant Answers and Significant Questions.” *Neuron*, vol. 70, no. 4, May 2011, pp. 687–702. PubMed Central, <https://doi.org/10.1016/j.neuron.2011.05.001>.
- Moreno-Jiménez, Elena P., et al. “Evidence for Adult Hippocampal Neurogenesis in Humans.” *Journal of Neuroscience*, vol. 41, no. 12, Mar. 2021, pp. 2541–53. Dual Perspectives. www.jneurosci.org, <https://doi.org/10.1523/JNEUROSCI.0675-20.2020>.
- Mortada, Ibrahim, et al. “Immunotherapies for Neurodegenerative Diseases.” *Frontiers in Neurology*, vol. 12, June 2021. Frontiers, <https://doi.org/10.3389/fneur.2021.654739>.
- Müller-Nedebock, Amica C., et al. “Different Pieces of the Same Puzzle: A Multifaceted Perspective on the Complex Biological Basis of Parkinson’s Disease.” *Npj Parkinson’s Disease*, vol. 9, no. 1, July 2023, p. 110. www.nature.com, <https://doi.org/10.1038/s41531-023-00535-8>.
- Nottebohm, F. “From Bird Song to Neurogenesis.” *Scientific American*, vol. 260, no. 2, Feb. 1989, pp. 74–79. PubMed, <https://doi.org/10.1038/scientificamerican0289-74>.
- Nowell, Joseph, et al. “Antidiabetic Agents as a Novel Treatment for Alzheimer’s and Parkinson’s Disease.” *Ageing Research Reviews*, vol. 89, Aug. 2023, p. 101979. ScienceDirect, <https://doi.org/10.1016/j.arr.2023.101979>.
- Obernier, Kirsten, and Arturo Alvarez-Buylla. “Neural Stem Cells: Origin, Heterogeneity and Regulation in the Adult Mammalian Brain.” *Development (Cambridge, England)*, vol. 146, no. 4, Feb. 2019, p. dev156059. PubMed Central, <https://doi.org/10.1242/dev.156059>.
- “Oxidation-Reduction Reactions.” Chemistry LibreTexts, 2 Oct. 2013, [https://chem.libretexts.org/Bookshelves/Analytical_Chemistry/Supplemental_Modules_\(Analytical_Chemistry\)/Electrochemistry/Redox_Chemistry/Oxidation-Reduction_Reactions](https://chem.libretexts.org/Bookshelves/Analytical_Chemistry/Supplemental_Modules_(Analytical_Chemistry)/Electrochemistry/Redox_Chemistry/Oxidation-Reduction_Reactions).
- Palanisamy, Chella Perumal, et al. “New Strategies of Neurodegenerative Disease Treatment with Extracellular Vesicles (EVs) Derived from Mesenchymal Stem Cells (MSCs).” *Theranostics*, vol. 13, no. 12, July 2023, pp. 4138–65. PubMed Central, <https://doi.org/10.7150/thno.83066>.
- Parent, Martin, et al. “Dopaminergic Innervation of the Human Subventricular Zone: A Comparison between Huntington’s Chorea and Parkinson’s Disease.” *American Journal of Neurodegenerative Disease*, vol. 2, no. 3, Sept. 2013, pp. 221–27.
- Parts of the Brain: Neuroanatomy, Structure & Functions in Neuroscience. 12 May 2025, <https://www.simplypsychology.org/anatomy-of-the-brain.html>. Biopsychology.

- Peretto, Paolo, and Luca Bonfanti. “Major Unsolved Points in Adult Neurogenesis: Doors Open on a Translational Future?” *Frontiers in Neuroscience*, vol. 8, June 2014, p. 154. PubMed Central, <https://doi.org/10.3389/fnins.2014.00154>.
- Perusini, Jennifer N., et al. “Optogenetic Stimulation of Dentate Gyrus Engrams Restores Memory in Alzheimer’s Disease Mice.” *Hippocampus*, vol. 27, no. 10, 2017, pp. 1110–22. Wiley Online Library, <https://doi.org/10.1002/hipo.22756>.
- Piscaglia, Anna Chiara. “Stem Cells, a Two-Edged Sword: Risks and Potentials of Regenerative Medicine.” *World Journal of Gastroenterology : WJG*, vol. 14, no. 27, July 2008, pp. 4273–79. PubMed Central, <https://doi.org/10.3748/wjg.14.4273>.
- Porlan, Eva, et al. “Paracrine Regulation of Neural Stem Cells in the Subependymal Zone.” *Archives of Biochemistry and Biophysics*, Factors controlling neural stem cells differentiation and renewal, vol. 534, no. 1, June 2013, pp. 11–19. ScienceDirect, <https://doi.org/10.1016/j.abb.2012.10.001>.
- Post-Transcriptional Control of Gene Expression | Biology for Majors I. <https://courses.lumenlearning.com/wm-biology1/chapter/reading-post-translational-control-of-gene-expression/>. Accessed 27 July 2025.
- Rajan, Arjun, et al. “Regulation of Neural Stem Cell Competency and Commitment during Indirect Neurogenesis.” *International Journal of Molecular Sciences*, vol. 22, no. 23, Nov. 2021, p. 12871. PubMed Central, <https://doi.org/10.3390/ijms222312871>.
- Ramón y Cajal, Santiago. *Degeneration and Regeneration of the Nervous System*. 1928. Google Scholar, <https://psycnet.apa.org/record/1928-10334-000>.
- Reynolds, B. A., and S. Weiss. “Generation of Neurons and Astrocytes from Isolated Cells of the Adult Mammalian Central Nervous System.” *Science (New York, N.Y.)*, vol. 255, no. 5052, Mar. 1992, pp. 1707–10. PubMed, <https://doi.org/10.1126/science.1553558>.
- Romito, Antonio, and Gilda Cobellis. “Pluripotent Stem Cells: Current Understanding and Future Directions.” *Stem Cells International*, vol. 2016, 2016, p. 9451492. PubMed Central, <https://doi.org/10.1155/2016/9451492>.
- Rozo, Jairo A., et al. “Cajal, the Neuronal Theory and the Idea of Brain Plasticity.” *Frontiers in Neuroanatomy*, vol. 18, Feb. 2024, p. 1331666. PubMed Central, <https://doi.org/10.3389/fnana.2024.1331666>.
- Ruetz, Tyson J., et al. “CRISPR–Cas9 Screens Reveal Regulators of Ageing in Neural Stem Cells.” *Nature*, vol. 634, no. 8036, Oct. 2024, pp. 1150–59. www.nature.com, <https://doi.org/10.1038/s41586-024-07972-2>.
- Sawai, Tsutomu, and Masanori Kataoka. “The Ethical and Legal Challenges of Human Foetal Brain Tissue-Derived Organoids.” *EMBO Reports*, vol. 25, no. 4, Mar. 2024, pp. 1700–03. PubMed Central, <https://doi.org/10.1038/s44319-024-00099-5>.
- Schacter, Daniel L., et al. “Specificity of Priming: A Cognitive Neuroscience Perspective.” *Nature Reviews Neuroscience*, vol. 5, no. 11, Nov. 2004, pp. 853–62. www.nature.com, <https://doi.org/10.1038/nrn1534>.
- Schieber, Michael, and Navdeep S. Chandel. “ROS Function in Redox Signaling and Oxidative

- Stress.” *Current Biology : CB*, vol. 24, no. 10, May 2014, pp. R453–62. PubMed Central, <https://doi.org/10.1016/j.cub.2014.03.034>.
- Schulte, Claudia, and Thomas Gasser. “Genetic Basis of Parkinson’s Disease: Inheritance, Penetrance, and Expression.” *The Application of Clinical Genetics*, vol. 4, June 2011, pp. 67–80. PubMed Central, <https://doi.org/10.2147/TACG.S11639>.
- Shah, Abdul Jalil, et al. “Neural Stem Cell Therapy for Alzheimer’s Disease: A-State-of-the-Art Review.” *Journal of Dementia and Alzheimer’s Disease*, vol. 1, no. 2, Dec. 2024, pp. 109–25. www.mdpi.com, <https://doi.org/10.3390/jdad1020008>.
- Stahnisch, Frank W. “A Century of Brain Regeneration Phenomena and Neuromorphological Research Advances, 1890s–1990s—Examining the Practical Implications of Theory Dynamics in Modern Biomedicine.” *Frontiers in Cell and Developmental Biology*, vol. 9, Jan. 2022, p. 787632. PubMed Central, <https://doi.org/10.3389/fcell.2021.787632>.
- Tay, Emmy Xue Yun, et al. “Epigenetic Plasticity and Redox Regulation of Neural Stem Cell State and Fate.” *Free Radical Biology and Medicine, Epigenetics and Redox Signaling*, vol. 170, July 2021, pp. 116–30. ScienceDirect, <https://doi.org/10.1016/j.freeradbiomed.2021.02.030>.
- Vassal, Mariana, et al. “Emerging Pro-Neurogenic Therapeutic Strategies for Neurodegenerative Diseases: A Review of Pre-Clinical and Clinical Research.” *Molecular Neurobiology*, vol. 62, no. 1, Jan. 2025, pp. 46–76. Springer Link, <https://doi.org/10.1007/s12035-024-04246-w>.
- Walgrave, Hannah, et al. “Restoring miR-132 Expression Rescues Adult Hippocampal Neurogenesis and Memory Deficits in Alzheimer’s Disease.” *Cell Stem Cell*, vol. 28, no. 10, Oct. 2021, pp. 1805-1821.e8. PubMed, <https://doi.org/10.1016/j.stem.2021.05.001>.
- “What Is Epigenetics?” Cleveland Clinic, <https://my.clevelandclinic.org/health/articles/epigenetics>. Accessed 24 July 2025.
- What Is Immunotherapy? | Immunotherapy for Cancer. <https://www.cancer.org/cancer/managing-cancer/treatment-types/immunotherapy.html>. Accessed 27 July 2025.
- Yen, Ya-Ping, and Jun-An Chen. “The m6A Epitranscriptome on Neural Development and Degeneration.” *Journal of Biomedical Science*, vol. 28, May 2021, p. 40. PubMed Central, <https://doi.org/10.1186/s12929-021-00734-6>.
- Zhao, Xinyu, and Darcie Moore. “Neural Stem Cells: Developmental Mechanisms and Disease Modeling.” *Cell and Tissue Research*, vol. 371, no. 1, Jan. 2018, pp. 1–6. PubMed Central, <https://doi.org/10.1007/s00441-017-2738-1>.
- Zou, Junhui, et al. “Methods to Measure Olfactory Behavior in Mice.” *Current Protocols in Toxicology / Editorial Board, Mahin D. Maines (Editor-in-Chief) ... [et Al.]*, vol. 63, Feb. 2015, p. 11.18.1-11.18.21. PubMed Central, <https://doi.org/10.1002/0471140856.tx1118s63>.

How Can Marketing Interventions Be Designed to Increase Attendance at High-School Athletic Events? By Roman Davis, Shemal Doshi

Background

Attendance at athletic events, such as soccer, volleyball, basketball, or football games, is typically high in large schools. However, small private schools often struggle with low turnouts. Despite this challenge, there is limited understanding of the effectiveness of marketing and promotional activities in persuading and motivating private school students and their families to attend these events. Developing targeted strategies to boost attendance remains an area in need of further exploration.

One study focused on a college basketball program and analyzed the mean effectiveness of different marketing and promotional activities in influencing and motivating college students to attend Division II athletic events. The study revealed that behavioral response incentives and marketing tactics that augment the sport product to better match the primary motive for fan attendance are the most effective. The opportunity to tailgate with friends and family at games was perceived as the most effective effort in augmenting attendance. However, the next most effective strategies were receiving free giveaways during the event and knowing that the athletic event supports a charitable organization. These results suggest that people are more likely to attend events when they feel they are getting something extra or valuable for free, but supporting a charitable organization is also valued.

In a study exploring the effectiveness of promotions used by BGSU athletics, BGSU students' motivations to attend college athletic events were compared with members of the Bowling Green Community. The results showed that there is not a significant difference between community members and BGSU students on the sub-factors of achievement, knowledge, aesthetics, drama, and escape. However, there is a significant difference between community members and BGSU students for the factors of social and family. Specifically, BGSU students were more motivated to attend sporting events because of the social context than community members. Meanwhile, community members were more motivated to attend sporting events because of family. These factors can be used to develop promotion strategies and guide the administration to promote the social aspect more to BGSU students and the family aspect to community members. Thus, it should be more effective if social promotions (e.g., having basketball tournaments or forms of competition in which students can participate with friends prior to games) are planned for student groups. Contrarily, some family promotions including hosting youth sport events that partner with BGSU athletics or providing family bundle ticket discounts would be effective for community members.

In another study, focused on Major League Baseball, researchers looked at how evolving family structures have affected marketing strategies. Standardized family packages such as deals for two adults and two children were offered, but the study revealed that these packages are no longer as effective as they once were. As modern families have become more diverse in size, makeup, and interests, the traditional "family of four" model has become no longer applicable to

many consumers. The study emphasized the need for organizations to adapt and customize their promotional strategies to fit the unique needs of today's families. This finding points to a larger trend in marketing: a shift away from "one-size-fits-all" methods in favor of more flexible, individualized approaches.

Despite these valuable insights, none of the above studies evaluated marketing interventions in the K-12 private school setting. While the college basketball study confirmed that both free giveaways and supporting a charitable organization are effective in attracting attendees, these two strategies were not directly compared to each other. Moreover, prior research has not addressed whether differences exist in attendance motivations between students and parents at the K-12 level. By identifying these gaps, our project aims to contribute new knowledge to the field and inform future promotional strategies for private schools.

This research project examines how various marketing activities affect attendance at athletic events in a small private K-12 school, specifically comparing free gifts to charitable donations. It also explores potential differences in marketing preferences by gender and age, seeking to provide actionable recommendations for school administrators.

Methods

An online questionnaire was distributed via email to all middle and high school students at The Meadows School (Las Vegas, NV), as well as their parents. The email, titled "*Survey Participation Request*," invited recipients to complete a brief survey aimed at identifying factors that influence attendance at school sporting events.

To ensure a comprehensive understanding of attendance drivers, the survey collected demographic information and assessed personal motivations. Demographic questions included:

- **Gender:** Female, Male, Non-binary/Genderqueer/Gender Fluid, Prefer not to say
- **Current Grade:** 6th–12th, Other (Parent)

Participants were then presented with three football game advertisements and asked to rate their willingness to attend after viewing each. The following instructions were given to the participants: You will see three different types of advertisement for the football game. Review each of them carefully and answer a question related to them. Each of the 3 advertisements included the following question: How willing are you to attend the event after seeing this advertisement?

The advertisements were identical except for an incentive statement added to two versions.

1. **Advertisement 1:** Plain football game advertisement (Figure 1).
2. **Advertisement 2:** Same advertisement with the statement: "*Show this email at the concession stand for free pizza!*" (Figure 2).

3. **Advertisement 3:** Same advertisement with the statement: “*Show this email at the concession stand and we will donate \$1 to the Leukemia & Lymphoma Society!*” (Figure 3).

Each response was recorded on a Numerical Rating Scale (NRS) from 1 (not at all interested) to 10 (absolutely committed and excited).

Additional questions included

4. **Which advertisement do you think would be most effective in increasing attendance?:** First, Second, Third
5. **Why do you think the option you chose would be most effective?:** Free-text response.



Figure 1: General advertisement



Figure 2: Advertisement with free pizza



Figure 3: Advertisement with charitable activity

Statistical Analysis

Descriptive statistics was used to summarize demographic characteristics. Paired t-tests were performed to compare willingness-to-attend ratings across the three advertisements to determine whether differences were statistically significant. Welch’s t-test was performed to test potential differences between males and females.

Results

A total of 517 middle and high school students, parents, faculty, and alumni at The Meadows School received the survey invitation. Of these, 455 recipients (88%) opened the email, and 212 (41%) completed the survey. Respondents represented a broad range of grades and age groups: 27.4% were middle school students, 50.9% high school students, and 21.7% parents (see Table 1 for demographic details). Gender distribution was balanced overall (42.9% male), though males were underrepresented among middle school respondents (19%).

Age Group	Number of Responders (Percent)				
	Total	Female	Male	Non-binary/ gender queer/ gender fluid	Prefer not to say
Total Population	212 (100%)	118 (55.7%)	91 (42.9%)	1 (0.5%)	2 (0.9%)
Middle School	58 (100%)	47 (81%)	11 (19%)	0	0
Upper School	108 (100%)	45 (41.7%)	60 (55.6%)	1 (0.9%)	2 (1.9%)
Parents	46 (100%)	26 (56.5%)	20 (43.5%)	0	0

Table 1: Demographics of the responders by age group and gender

Attendance Motivation by Advertisement Type

Free Pizza vs. General Advertisement The advertisement offering free pizza (Ad 2) received significantly higher mean ratings than the general advertisement (Ad 1) among the total population, middle school, and high school students, but not among adults (Table 2, Figure 4).

Charity vs. General Advertisement The charity-focused advertisement (Ad 3) was rated significantly higher than the general advertisement across all age groups.

Charity vs. Free Pizza: Middle school students and adults rated the charity advertisement higher than free pizza, while high school students preferred free pizza. Although the overall mean rating for the charity ad was numerically higher than free pizza, this difference was not statistically significant (p-value > 0.05).

Overall, both the free pizza and charity advertisements were significantly more effective than the general advertisement in increasing willingness to attend. High school students favored free pizza, whereas middle school students and adults leaned toward charitable promotions.

Age Group	Responses to Questions 1- 3: Numerical Rating Scale (NRS) (Scores 1 - 10)						Paired T-test		
	Q1. Rating of the general advertisement		Q2. Rating of the advertisement with free pizza		Q3. Rating of the advertisement with charity		Q1 vs Q2	Q1 vs Q3	Q2 vs Q3
	Median	Mean	Median	Mean	Median	Mean	p-value <0.05?		
Total Population N=212	5	4.70	7	6.75	7	6.88	Yes	Yes	No
Middle School N=58	5	4.55	7	6.76	8	7.53	Yes	Yes	Yes
Upper School N=108	4	4.51	8	7.06	7	6.50	Yes	Yes	Yes
Parents N=46	5	5.35	6.5	6.00	8	6.93	No	Yes	Yes

Table 2: Mean and median NRS scores for each advertisement (with paired t-test results) by age group.

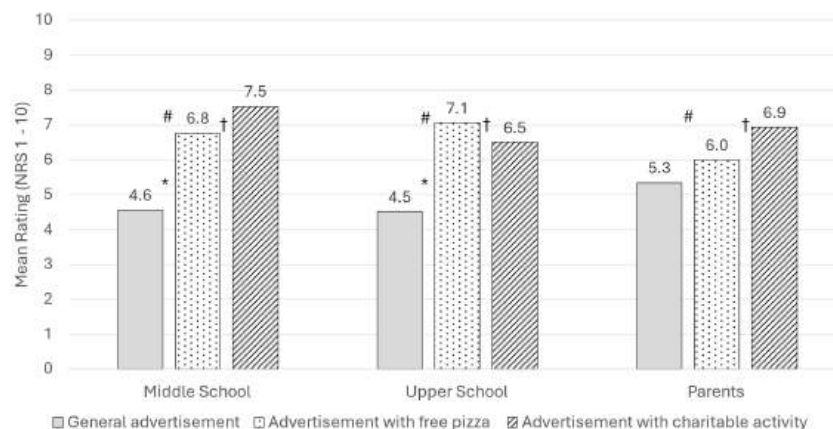


Figure 4: Mean willingness to attend by advertisement type and age group.

(*) indicate statistically significant differences (p-value < 0.05) between general advertisement and advertisement with free pizza

(#) indicate statistically significant differences (p-value < 0.05) between general advertisement and advertisement with charitable activity

(†) indicate statistically significant differences (p-value < 0.05) between advertisement with free pizza and advertisement with charitable activity

Although male respondents ranked all three ads higher, the difference between groups was not significant (Table 3). The analysis of the responses to Questions 1 - 3 by gender revealed differences in preference between males and females, particularly in the middle school and adult age groups (Table 3, Figure 5). Females preferred the charity advertisement and males overall had stronger preference for free pizza, but this was primarily driven by high school males. In fact, male middle-school and adult population had numerically higher ratings for charity than free food.

Gender/Age Group	Responses to Questions 1- 3: Numerical Rating Scale (NRS) (Scores 1 - 10)						Paired T-test		
	Q1. Rating of the general advertisement		Q2. Rating of the advertisement with free pizza		Q3. Rating of the advertisement with charity		Q1 vs Q2	Q1 vs Q3	Q2 vs Q3
	Median	Mean	Median	Mean	Median	Mean	p-value <0.05?		
Female Responders N=118	5	4.44	7	6.29	7	6.86	Yes	Yes	Yes
Female Middle School Responders N=47	5	4.36	7	6.66	8	7.51	Yes	Yes	Yes
Female Upper School Responders N=45	4	4.04	6	6.22	6	6.11	Yes	Yes	No
Female Adult Responders N=26	5	5.27	6	5.73	8	6.96	No	Yes	No
Male Responders N=91	5	5.13	8	7.32	7	6.93	Yes	Yes	No
Male Middle School Responders N=11	5	5.36	5	7.18	5	7.64	Yes	Yes	No
Male Upper School Responders N=60	5	4.98	8	7.67	7	6.82	Yes	Yes	Yes
Male Adult Responders N=20	5	5.45	5	6.35	7.5	6.9	No	Yes	No

Table 3: Mean and median NRS scores for each advertisement (with paired t-test results) by age group and gender.

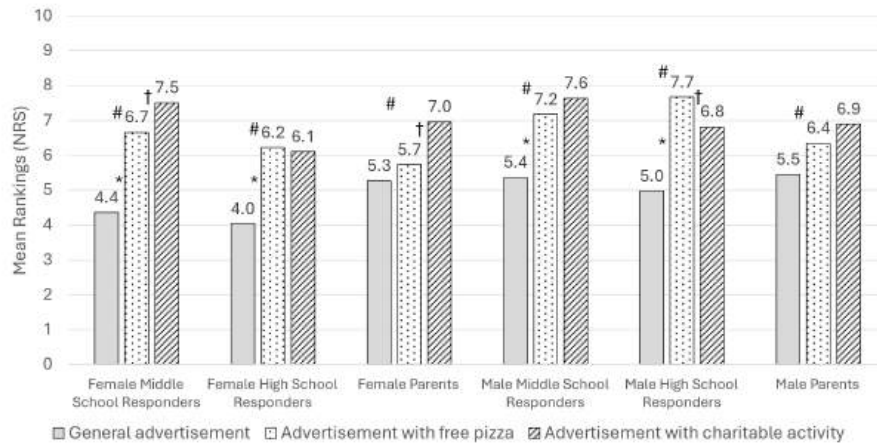


Figure 5: Mean NRS ratings for each advertisement by age group and gender.

(*) indicate statistically significant differences (p-value < 0.05) between general advertisement and advertisement with free pizza

(#) indicate statistically significant differences (p-value < 0.05) between general advertisement and advertisement with charitable activity

(†) indicate statistically significant differences (p-value < 0.05) between advertisement with free pizza and advertisement with charitable activity.

Perceived Effectiveness of Advertisements

Despite the higher overall rankings for the charity advertisement, the majority of respondents in all age groups selected the free pizza advertisement as the most effective for increasing attendance (Table 3, Figure 5). The only subgroup with a slight preference for charity over free pizza was male parents. Free-text responses largely confirmed these selections, with many respondents citing the universal appeal of free food.

Age Group	Number of Responders (Percent)					
	Q4: Which advertisement do you think would be most effective in getting a higher number of attendees?			Responses to Q5: Why do you think the option you chose would be the most effective? (free text)		
	Ad #1 (general)	Ad #2 (free food)	Ad #3 (charity)	Free text confirming selection in Q4	Free text contradicting selection in Q4	Free text not explaining selection in Q4
Total Population N=212	4 (1.9%)	140 (66%)	68 (32.1%)	199 (93.9%)	6 (2.8%)	7 (3.3%)
Middle School N=58	1 (1.7%)	36 (62.1%)	21 (36.2%)	50 (86.2%)	5 (8.6%)	3 (5.2%)
Upper School N=108	2 (1.9%)	79 (73.1%)	27 (25%)	104 (96.3%)	1 (0.9%)	3 (2.8%)
Parents/Faculty N=46	1 (2.2%)	25 (54.3%)	20 (43.5%)	45 (97.8%)	0	1 (2.2%)

Table 4: Perceived effectiveness of each advertisement and free text rationale by age group.

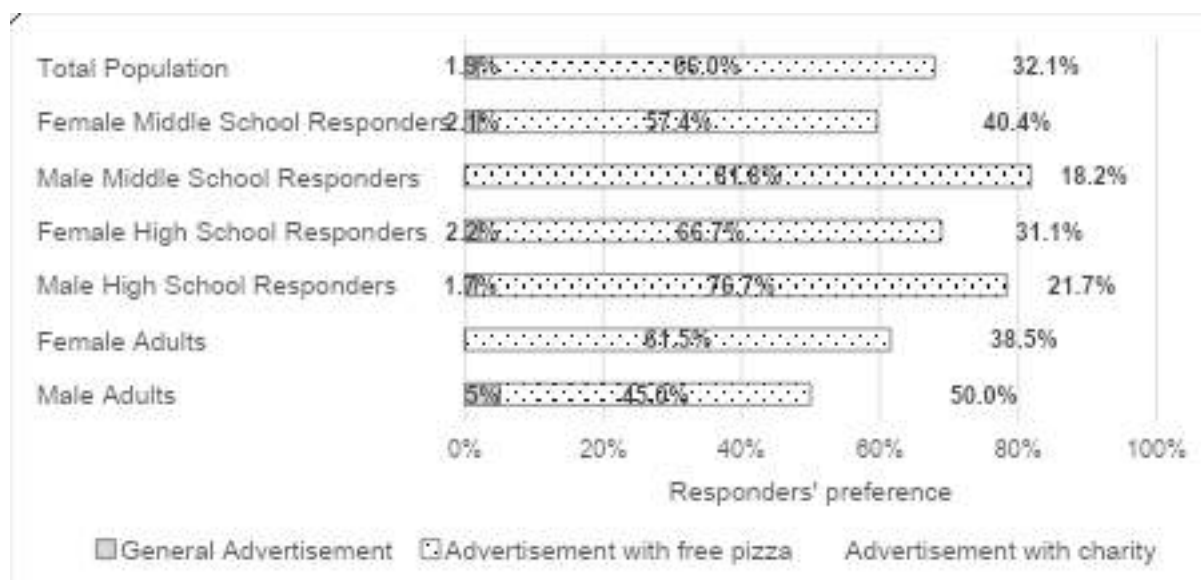


Figure 5: Percentage of respondents selecting each advertisement as most effective, by age group and gender.

Key Findings

Both free giveaways and charitable promotions significantly increased willingness to attend compared to general advertisements.

High school students were most motivated by free food, while middle school students and adults responded more to charitable incentives.

The majority believed free food would be most effective in boosting attendance, despite higher ratings for charity ads among some groups.

Discussion

This study addresses a critical challenge of low attendance at athletic events faced by small private schools. While previous research has focused on college and professional sports settings, our work provides new valuable insights into the effectiveness of marketing interventions in a K–12 private school environment, specifically comparing the impact of free giveaways and charitable promotions.

Our results demonstrate that both free food incentives and charitable promotions significantly increase students' and parents' willingness to attend athletic events compared to general advertisements. We also found that preferences varied by age group and gender: high school students were most motivated by free food, whereas middle school students and adults showed a stronger response to charitable incentives. Despite higher mean ratings for charity advertisements among some groups, the majority of respondents believed that free food would be most effective in boosting attendance. This finding was confirmed by free-text responses.

The results of our research align with prior studies in collegiate and professional sports, which have shown that tangible incentives and social motivations can drive attendance. However, our study extends these findings to the K–12 private school context and highlights the importance of tailoring marketing strategies to specific demographic groups.

The implications for schools and organizations from our work are that marketing strategies should be customized to the preferences of different age groups and genders. For example, offering free food may be particularly effective for high school students, while charitable promotions may resonate more with middle school students and parents. By leveraging these insights, schools can potentially enhance community engagement and support of their athletic programs.

A major strength of this study is its focus on a small private school setting, filling a gap in the literature. The use of both quantitative ratings and qualitative free-text responses allowed for validation of survey engagement and provided richer context for interpreting the results. However, the study is limited by its reliance on self-reported data, which may introduce bias, and by its focus on a single school, which may limit generalizability, particularly since not all recipients completed the survey. Additionally, the cross-sectional design does not capture long-term effects of marketing interventions.

Future research should explore these strategies in diverse educational settings and with larger sample sizes to validate and extend our findings. Longitudinal studies could assess the sustained impact of different marketing approaches on event attendance. Further investigation into additional factors, such as event type, timing, and the role of peer influence, may also yield valuable insights for optimizing promotional efforts.

Conclusion

By understanding and leveraging the preferences of different demographic groups, schools and organizations can develop more effective marketing strategies that not only increase attendance but also foster a stronger sense of community. This research contributes to the growing body of knowledge on marketing interventions in educational settings and offers practical recommendations for enhancing event participation.

Works Cited

- 1.Barrett M. and Farr J.R. Marketing Division II athletics to college students: The perceived effectiveness of internally focused promotion tactics. *The Sport Journal*, 2023, Vol. 24.
- 2.Martinson D. et al. An Analysis of the Factors and Marketing Techniques Affecting Attendance at NCAA Division I Women's Basketball Games. *The Journal of SPORT*, 2015, 4(2), 42-59
- 3.Cianfrone BA, et al. Identifying Key Market Demand Factors Associated with High School
- 4.Basketball Tournaments. *Sport Marketing Quarterly*, 2015, 24, 91-104.
- 5.Sport Organizations Frequently Strike Out When Marketing to Families. College of Business, July 17, 2024:
<https://www.wcu.edu/learn/departments-schools-colleges/cob/about-the-college-of-business/ideas-with-impact/posts/sport-organizations-frequently-strike-out.aspx?utm>
- 6.Browning, R. A. & DeBolt, L. S. The effects of promotions on attendance in professional baseball. *The Sport Journal*, 2007: 10(3).
- 7.Choi, Y., et al. Motivational factors influencing sport spectator involvement at NCAA Division II basketball games. *Journal for the Study of Sports and Athletes in Education*, 2009: 3(3), 265-284.
- 8.Smith A.C.T. and Stewart B. *Introduction to sport marketing*. 2nd ed, Routledge, 2015.

Wick Rotation and Applications By Atharv Srivastava

Abstract

We examine Wick rotation as an analytic continuation of time and its applications in curved space times. Beginning with the distinction between Lorentzian and Euclidean metrics and the role of curvature, we present Wick rotation as a complex transformation that alters causal structure and metric signature. Limitations arising from the absence of global timelike Killing vectors are discussed with emphasis on Schwarzschild and Friedmann–Robertson–Walker (FRW) spacetimes, where only local continuations are generally valid. Applications are then examined: in black hole physics, Wick rotation leads to the emergence of periodic imaginary time, providing a natural derivation of the Bekenstein–Hawking entropy via conical defects; in analog gravity systems such as phonons in fluids and Bose–Einstein condensates, Wick rotation provides a powerful framework to analyze thermal effects near horizons. Altogether, the study demonstrates how Wick rotation serves as a unifying tool, connecting analytic continuation, spacetime geometry, black hole thermodynamics, and laboratory analogs of gravity.

1. Metrics

1.1 Pythagoras to Differential geometry

A metric essentially defines distance, angle, and geometry. The very idea of a *metric* is to provide a rule for measuring the separation between infinitesimally close events. In an n -dimensional smooth manifold M , a (pseudo-)Riemannian metric is a symmetric, non-degenerate bilinear form g acting on tangent vectors:

$$ds^2 = g_{\mu\nu}(x) dx^\mu dx^\nu, \quad (1)$$

written here in local coordinates $\{x^\mu\}$ (Greek indices run $0, \dots, n-1$ and we adopt the Einstein summation convention).

Matrix representation. At each point p in the manifold M , $g_{\mu\nu}(p)$ may be viewed as an $n \times n$ matrix whose *signature*¹⁵ controls the local causal structure. Line element (1) is *invariant* under coordinate changes, reflecting the geometric—not merely algebraic—character of the metric.

1.2 Lorentzian vs Euclidean

- Lorentzian metrics have signatures $(-, +, \dots, +)$. The archetype is the Minkowski metric, which describes the geometry of special relativity $\eta_{\mu\nu} = \text{diag}(-1, 1, 1, 1)$. Here, the ‘-’ signed time coordinate signifies a complex time axis.
- Euclidean metrics have an all-positive signature $(+, +, \dots, +)$; A trivial example would be the flat $\mathbb{R}^4$ space, with the metric $\delta_{ij} = \text{diag}(1, 1, 1, 1)$.

¹⁵ The signature counts $(-)$ and $(+)$ eigenvalues.

2. Curvature

Intuitively, curvature quantifies how much the geometry deviates from being flattenable. Lines and angles differ in curved space when compared to flat space. Curvature can be interpreted in multiple ways. The angle sum property of triangles does not hold true in geometries with positive/negative curvature (spherical/hyperbolic).

From a differential geometry point of view, we study how vectors translate along the space, also called parallel transport. The Levi-Civita connection, $\Gamma_{\mu\nu}^{\rho} = \frac{1}{2}g^{\rho\sigma}(\partial_{\mu}g_{\nu\sigma} + \partial_{\nu}g_{\mu\sigma} - \partial_{\sigma}g_{\mu\nu})$, captures how vectors parallel-transport. Curvature enters through the Riemann tensor,

$$R_{\sigma\mu\nu}^{\rho} = \partial_{\mu}\Gamma_{\nu\sigma}^{\rho} - \partial_{\nu}\Gamma_{\mu\sigma}^{\rho} + \Gamma_{\mu\lambda}^{\rho}\Gamma_{\nu\sigma}^{\lambda} - \Gamma_{\nu\lambda}^{\rho}\Gamma_{\mu\sigma}^{\lambda}, \quad (2)$$

with contractions giving the Ricci tensor $R_{\mu\nu}$ and scalar R .

2.1 The Schwarzschild Metric

Outside a static, spherically symmetric mass M , general relativity predicts

$$ds^2 = -\left(1 - \frac{2GM}{r}\right)c^2dt^2 + \left(1 - \frac{2GM}{r}\right)^{-1}dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2). \quad (3)$$

Here $r_s = 2GM$ is the Schwarzschild radius. For the cases $r = 0$ and $r = r_s$, the coefficients of dt^2 and dr^2 diverge, giving us two 'singularities', the former referring to the spacetime singularity at the center of the black hole, while the latter refers to the boundary known as the *event horizon*. Interestingly, the region $0 < r < r_s$ shows a peculiar behavior, wherein space and time seem to switch roles. As we cross the event horizon, instead of being able to move only in one direction in time (forwards), we end up having to move only in one direction in space, i.e. towards the singularity. This means that time does not show a global behavior, something that we will come back to, later in this paper.

3. Wick Rotation: Analytic Continuation in Time

3.1 Why Call It a "Rotation"?

In complex analysis, multiplying a real variable by i rotates the real line into the imaginary axis. Setting

$$t \mapsto -i, \quad \in \mathbb{R}, \quad (4)$$

turns the Lorentzian interval $ds^2 = -c^2dt^2 + d\vec{x}^2$ into the Euclidean norm

$$ds_E^2 = c^2d^2 + d\vec{x}^2 \quad \text{a genuine rotation in the } t\text{-}it \text{ plane.}$$

3.2 Analytic Continuation: Theory and Examples

In complex analysis, a function f is *analytic* on a domain $U \subset \mathbb{C}$ if it is locally given by a convergent power series. Often we first meet f through a series or integral that only *initially* converges on a smaller set $U_0 \subset U$. *Analytic continuation* is the process of extending f to a larger

domain U in such a way that the extension is still analytic and agrees with the original f on U_0 .

3.3 Geometric Interpretation

The causal structure encoded by light cones disappears: null rays ($ds^2 = 0$) in Lorentzian signatures become *points* in Euclidean space.

4 Limitations of Wick Rotation

4.1 Global Obstructions

A global Wick rotation demands:

- (a) a *globally defined time function* t ,
- (b) a nowhere-vanishing, timelike Killing vector ∂_t ,
- (c) absence of closed timelike curves (CTCs).

Spacetimes lacking these features (Godel, Kerr interior, certain cosmologies) resist a neat analytic continuation.

4.2 Killing Vectors and Killing Coordinates

Symmetries of the metric. A smooth vector field K^μ on a spacetime (M, g) is called a *Killing vector* if it generates an isometry, i.e. the Lie derivative of the metric vanishes:

$$\mathcal{L}_K g_{\mu\nu} = K^\rho \nabla_\rho g_{\mu\nu} + (\nabla_\mu K^\rho) g_{\rho\nu} + (\nabla_\nu K^\rho) g_{\rho\mu} = 0. \quad (5)$$

In components, $\nabla_{(\mu} K_{\nu)} = 0$, which is just the infinitesimal statement that g is unchanged when we flow along K .

Adapted (“Killing”) coordinates. Whenever a Killing vector K is *nowhere vanishing* on an open set, one can choose local coordinates $x^0, x^1, \dots, x^{n-1}$ so that

$$K^\mu \partial_\mu = \partial_{x^0} \quad \text{and} \quad \partial_{x^0} g_{\alpha\beta} = 0. \quad (6)$$

That is, the metric components are *independent* of x^0 : the corresponding coordinate plays the rôle of a symmetry direction and is called a *Killing coordinate* (sometimes “ignorable” or “cyclic” coordinate).

Timelike Killing vectors and stationarity. If K^μ is everywhere timelike ($g_{\mu\nu} K^\mu K^\nu < 0$), the spacetime is said to be *stationary*. A stronger condition—*static* spacetime—also requires that K be hypersurface orthogonal: $K_{[\mu} \nabla_\nu K_{\rho]} = 0$. In either case, the integral curves of K provide a natural global time parameter, often denoted t , whose flow leaves $g_{\mu\nu}$ unchanged.

Why do we need a timelike Killing coordinate for Wick rotation?

- (a) Analytic continuation target. Wick rotation amounts to $t \mapsto -i$ in *one specific*

coordinate. For the resulting Euclidean metric to inherit smoothness and (ideally) real coefficients, t must already be a coordinate in which $g_{\mu\nu}$ is *independent of* t . Equation (6) supplies exactly that guarantee.

(b) Global coverage. If K fails to be timelike (or vanishes) somewhere, t ceases to be globally valid; the rotated metric may become complex or singular there. Schwarzschild exemplifies this: $K = \partial_t$ is timelike only for $r > 2GM$, so the Euclidean extension lives *outside* the horizon and regularity enforces a periodic imaginary time.

When Killing vectors are absent. Cosmological spacetimes such as FLRW possess no timelike Killing vector (the universe is expanding), so there is no single global coordinate that satisfies (6). One may still perform *local* Wick rotations patch by patch, but the resulting Euclidean manifolds must be sewn together with care, and the physical interpretation becomes subtler.

Illustrative examples.

- **Minkowski:** $K = \partial_t$ is globally timelike, giving the classic $t \rightarrow -i$ with a flat $\mathbb{R}^4$ Euclidean space.
- **Schwarzschild (exterior):** $K = \partial_t$ is timelike only for $r > 2GM$, so the Euclidean section is restricted to that region and is periodic with $\beta = 8\pi GM/c^3$.
- **Kerr:** There is a stationary Killing vector $K = \partial_t + \Omega_H \partial_\phi$ that is timelike near infinity but spacelike inside the ergosphere, thwarting a global rotation.

In short, the existence of a *globally timelike Killing coordinate* provides the rigid scaffolding that lets us turn Lorentzian time into Euclidean (“imaginary”) time without tearing the manifold or introducing complex-valued metric coefficients. When that scaffolding is missing or incomplete, we must either restrict to patches, accept a periodic imaginary time with singularities excised, or abandon Wick rotation altogether in favour of real-time techniques.

4.3 Schwarzschild Revisited

Applying $t \rightarrow -i$ to (3) yields

$$ds_E^2 = \left(1 - \frac{2GM}{r}\right) c^2 d^2 + \left(1 - \frac{2GM}{r}\right)^{-1} dr^2 + r^2 d\Omega^2.$$

Demanding regularity at $r = r_s$ forces to be periodic with period $\beta = 8\pi GM/c^3$, beautifully reproducing the Hawking temperature $T_H = 1/\beta$. But note: the rotation works only *outside* the horizon and after extending t into the complex plane; globally, the Schwarzschild manifold fails condition (a).

5 Wick Rotation of the FRW Metric

The Friedmann–Robertson–Walker (FRW) line element in Lorentzian signature is

$$ds^2 = -dt^2 + a^2(t) \left[\frac{dr^2}{1 - kr^2} + r^2 d\Omega^2 \right], \quad (7)$$

where $a(t)$ is the scale factor and $k = 0, \pm 1$ specifies the spatial curvature.

$$t \mapsto -i\tau, \quad dt \mapsto -id\tau, \quad (8)$$

A Wick rotation analytically continues the time coordinate into the complex plane, so that the Lorentzian line element becomes

$$ds_E^2 = d\tau^2 + a^2(-i\tau) \left[\frac{dr^2}{1 - kr^2} + r^2 d\Omega^2 \right]. \quad (9)$$

Observation tells us that the metric collapses to a singularity for $a(t) = 0$, i.e. the Big bang singularity. Hence, the analytic continuation $a(t) \rightarrow a(-i\tau)$ may not be regular at that point. In FRW spacetimes with horizons (e.g., de Sitter, open FRW with inflation), the Lorentzian manifold has regions that are not covered by a single global time coordinate t . Also, For arbitrary forms of $a(t)$, $a(-i\tau)$ is not necessarily well-defined or analytic for all t .

5.1 Local Wick Rotation

The solution to the above mentioned issues lies in the idea of local Wick rotation. If we consider a small slice of the spacetime, wherein the scale factor $a(t)$ can be assumed to be a constant a_0 , the resulting metric becomes

$$ds^2 = -dt^2 + a_0^2 \left[\frac{dr^2}{1 - kr^2} + r^2 d\Omega^2 \right] \quad (10)$$

On Wick rotation, we obtain the following form

$$ds_E^2 = +d\tau^2 + a_0^2 \left[\frac{dr^2}{1 - kr^2} + r^2 d\Omega^2 \right] \quad (11)$$

Hence, we have successfully rotated a Lorentzian metric into a Euclidean 4-sphere.

6. Using Wick Rotations in Calculating Black-Hole Entropy and its implications in Phonon Physics

6.1 Conventions

We work on a smooth, time-orientable Lorentzian manifold (M, g) of dimension $d \geq 3$ with signature $(- + \cdots +)$. Units $c = \hbar = k_B = 1$. $(\cdot)_E$ denotes Euclidean quantities obtained by the analytic continuation $t \mapsto -i\tau$.

6.2 Definitions

A vector field K^μ is *Killing* if $L_K g_{\mu\nu} = 0$. A *Killing horizon* $H \subset M$ is a null hypersurface whose null generators are tangent to a Killing field K^μ , and on which $K^2 \equiv g_{\mu\nu} K^\mu K^\nu = 0$. Assume H

is *non-extremal*: the *surface gravity* $\kappa > 0$, defined on H by the equivalent relations

$$K^\nu \nabla_\nu K^\mu = \kappa K^\mu, \quad \nabla_\mu (K^2) = -2\kappa K_\mu, \quad (12)$$

is strictly positive. Let Σ be a spacelike hypersurface intersecting H in a compact, smooth $(d-2)$ -manifold $B := H \cap \Sigma$ (the *horizon cross-section*). We call A_B the area of B .

6.3 Near-horizon normal form.

For notational simplicity we first treat the *static* case, in which K is hypersurface-orthogonal and one may write locally

$$ds^2 = -N^2(x) dt^2 + h_{ij}(x) dx^i dx^j, \quad g_{tt} = 0. \quad (13)$$

Choose Gaussian normal coordinates (ρ, y^A) on Σ near B , where $\rho \geq 0$ is proper distance orthogonal to B and y^A are coordinates along B :

$$h_{ij} dx^i dx^j = d\rho^2 + \gamma_{AB}(\rho, y) dy^A dy^B, \quad \gamma_{AB}(0, y) = \gamma_{AB}^{(0)}(y). \quad (14)$$

Non-extremality implies that the lapse vanishes linearly,

$$N(\rho, y) = \kappa \rho + O(\rho^2), \quad (15)$$

with constant κ on B (the *zeroth law*). Substituting (14)–(15) into (13) yields the *near-horizon normal form*

$$ds^2 = -\kappa^2 \rho^2 dt^2 + d\rho^2 + \gamma_{AB}^{(0)}(y) dy^A dy^B + O(\rho^2). \quad (16)$$

The (t, ρ) sector is the 2D Rindler metric; the tensor $\gamma^{(0)}$ is the intrinsic metric on B . 5

6.4 Euclideanization and the cone parameter.

After the Wick rotation $t = -i\tau$, (16) becomes

$$ds_E^2 = \kappa^2 \rho^2 d\tau^2 + d\rho^2 + \gamma_{AB}^{(0)}(y) dy^A dy^B + O(\rho^2). \quad (17)$$

Let $\beta > 0$ be the period of τ ($\tau \sim \tau + \beta$) and define the angular variable $\theta \sim \theta + \kappa\beta$. Near $\rho = 0$ the 2D (ρ, θ) geometry is that of a cone:

$$ds_{(2)}^2 = d\rho^2 + \alpha^2 \rho^2 d\theta^2, \quad \theta \in [0, 2\pi), \quad \alpha = \frac{\kappa\beta}{2\pi}. \quad (18)$$

The Euclidean section is smooth at $\rho = 0$ iff $\alpha = 1$ (equivalently, $\beta = 2\pi/\kappa$). For $\alpha \neq 1$ there is a conical singularity with deficit angle $2\pi(1 - \alpha)$. In Euclidean language, B is the fixed-point set (“bolt”) of the $U(1)$ generated by ∂_τ .

6.5 Curvature of a 2D cone.

Let $D_\epsilon = \{\rho \leq \epsilon\}$ be a small disk around $\rho = 0$ in the cone (18). By Gauss–Bonnet,

$$\int_{D_\epsilon} K dA + \int_{\partial D_\epsilon} k_g ds = 2\pi, \quad (19)$$

with Gaussian curvature K , geodesic curvature k_g of the boundary, and area form dA . Since ∂D_ϵ has length $L = 2\pi\alpha\epsilon$ and $k_g = 1/\epsilon$, one finds $\int_{\partial D_\epsilon} k_g ds = 2\pi\alpha$, hence

$$\int_{D_\epsilon} K dA = 2\pi(1 - \alpha). \quad (20)$$

Because $R_{(2)} = 2K$ in 2D, this yields the distributional identity

$$\int_{D_\epsilon} \sqrt{g_{(2)}} R_{(2)} = 4\pi(1 - \alpha) \implies R_{(2)} = 4\pi(1 - \alpha) \delta^{(2)}(\rho) \quad (21)$$

in the sense of integrals against smooth test functions.

6.6 Singular curvature in d dimensions

In a small tubular neighborhood $U_\epsilon \cong D_\epsilon \times \mathcal{B}$ of the bolt, the Euclidean metric factorises as (cone) $\times \mathcal{B}$ up to $O(\rho^2)$. The only singular contribution to the d -dimensional scalar curvature R_E comes from the cone factor. Integrating (21) over U_ϵ gives

$$\int_{U_\epsilon} \sqrt{g_E} R_E = 4\pi(1 - \alpha) A_{\mathcal{B}}, \quad A_{\mathcal{B}} \int_{\mathcal{B}} \sqrt{\det \gamma^{(0)}(y)} d^{d-2}y. \quad (22)$$

6.7 Euclidean action and its singular part

Consider the Euclidean gravitational action

$$I_E(\beta) = -\frac{1}{16\pi G} \int_{\mathcal{M}_E} \sqrt{g_E} R_E - \frac{1}{8\pi G} \int_{\partial \mathcal{M}_E} \sqrt{h} K + I_E^{(\text{matter})}, \quad (23)$$

with the usual Gibbons–Hawking–York term on the (outer) boundary ∂M_E . When one varies the Euclidean time period β *off shell*, holding the fields fixed away from the bolt, all terms in (23) except the curvature singularity scale linearly with β in a neighborhood of the smooth value $\beta_H 2\pi/\kappa$. Separating the contribution (22), one may write

$$I_E(\beta) = I_E^{\text{reg}}(\beta) - \frac{A_{\mathcal{B}}}{4G} (1 - \alpha(\beta)), \quad \alpha(\beta) = \frac{\kappa\beta}{2\pi}, \quad (24)$$

where I_E^{reg} is smooth and satisfies $I_E^{\text{reg}}(\beta) = C\beta + O((\beta - \beta_H)^2)$ for some constant C determined by the regular bulk and boundary data.

6.8 Entropy from the conical defect.

Interpreting $Z(\beta) \approx e^{-I_E(\beta)}$ as the canonical partition function, the free energy is $F = I_E/\beta$ and the thermodynamic entropy is

$$S = \beta (E - F) = \left(\beta \frac{\partial}{\partial \beta} - 1 \right) I_E(\beta) \Big|_{\beta=\beta_H}, \quad E = \frac{\partial I_E}{\partial \beta}. \quad (25)$$

Using (24) in (25) and noting that $(\beta \partial_\beta - 1) I_E^{\text{reg}}(\beta_H) = 0$ (because I^{reg} is linear in β at β_H), while $\partial_\beta \alpha = \kappa/(2\pi) = 1/\beta_H$, one obtains

$$S = \frac{A_{\mathcal{B}}}{4G} [\beta \partial_\beta \alpha(\beta) - (1 - \alpha(\beta))]_{\beta=\beta_H} = \frac{A_{\mathcal{B}}}{4G}. \quad (26)$$

Thus the Bekenstein–Hawking area law is recovered purely from the local cone structure near a non-extremal Killing horizon.

6.9 One concrete check: Schwarzschild

Consider the Schwarzschild metric in units $c = \hbar = k_B = 1$:

$$ds^2 = -f(r) dt^2 + \frac{dr^2}{f(r)} + r^2 d\Omega^2, \quad f(r) = 1 - \frac{2GM}{r}. \quad (27)$$

The (outer) horizon is at $r = r_h \equiv 2GM$, where $f(r_h) = 0$. The surface gravity is

$$\kappa = \frac{1}{2} f'(r_h) = \frac{1}{2} \left(\frac{2GM}{r^2} \right) \Big|_{r=2GM} = \frac{1}{4GM}. \quad (28)$$

To exhibit the near-horizon normal form, set $r = r_h + \varepsilon$ with $|\varepsilon| \ll r_h$. Then $f(r) = f'(r_h) \varepsilon + O(\varepsilon^2)$.

Define a proper radial coordinate ρ by

$$\varepsilon = \frac{1}{4} f'(r_h) \rho^2 = \frac{\rho^2}{8GM}, \quad \Rightarrow \quad d\varepsilon = \frac{1}{2} f'(r_h) \rho d\rho. \quad (29)$$

A short computation yields, to leading order in ρ ,

$$-f(r) dt^2 + \frac{dr^2}{f(r)} = -\frac{f'(r_h)^2}{4} \rho^2 dt^2 + d\rho^2 = -\kappa^2 \rho^2 dt^2 + d\rho^2, \quad (30)$$

so that the full metric takes the Rindler $\times$ sphere form

$$ds^2 = -\kappa^2 \rho^2 dt^2 + d\rho^2 + r_h^2 d\Omega^2 + O(\rho^2). \quad (31)$$

After Wick rotation $t = -i\tau$

$$ds_E^2 = \kappa^2 \rho^2 d\tau^2 + d\rho^2 + r_h^2 d\Omega^2 + O(\rho^2). \quad (32)$$

Regularity at $\rho = 0$ (absence of a conical defect) requires that τ be periodic with

$$\beta = \frac{2\pi}{\kappa} = 8\pi GM, \quad \Longleftrightarrow \quad T_H = \frac{1}{\beta} = \frac{1}{8\pi GM}. \quad (33)$$

The horizon (bolt) area is

$$A_B = 4\pi r_h^2 = 16\pi G^2 M^2, \quad (34)$$

and the conical–defect (or equivalently Wald/Euclidean) argument gives the Bekenstein–Hawking entropy

$$S_{\text{BH}} = \frac{A_B}{4G} = 4\pi G M^2. \quad (35)$$

6.10 Connection to analog gravity and usefulness.

In analog gravity, linear sound (phonon) excitations in a steady, irrotational fluid or Bose–Einstein condensate (BEC) propagate on an *acoustic metric*

$$ds_{\text{ac}}^2 \propto -(c^2 - v^2) dt^2 - 2\mathbf{v} \cdot d\mathbf{x} dt + d\mathbf{x}^2,$$

where $c(x)$ is the local sound speed and $\mathbf{v}(x)$ the background flow. A (sonic) horizon occurs where the normal component of the flow matches the sound speed, $|\mathbf{v}_\perp| = c$; the Killing generator is the laboratory time translation in a steady flow. Near such a horizon the metric reduces locally to the same Rindler form as above, with an “acoustic surface gravity” κ_{ac} defined by the normal gradient of the lapse $N_{\text{ac}} \equiv \sqrt{c^2 - v^2}$:

$$N_{\text{ac}}(\rho) = \kappa_{\text{ac}} \rho + O(\rho^2), \quad \Rightarrow \quad T_{\text{Hawking}}^{(\text{ac})} = \frac{\kappa_{\text{ac}}}{2\pi}.$$

The Euclidean regularity/cone argument therefore carries over verbatim, providing:

- a clean derivation of the *analog Hawking temperature* used to design and interpret horizon experiments in BECs, water channels, optical media, and related platforms;
- a Euclidean (imaginary–time) framework for *phonon thermodynamics* and noise, enabling calculations of partition functions, free energies, and instanton rates (e.g. phase slips in ring BECs) in steady configurations;
- a conceptual bridge: whenever a global (laboratory) timelike Killing field exists, equilibrium and reflection positivity arguments apply to phonons just as in static black–hole exteriors.

Here “phonons” are the quantised normal modes of the medium (density/phase waves in superfluids/BECs, lattice vibrations in solids). They are the effective quantum fields that “feel” the acoustic geometry, and thus serve as experimentally accessible probes of horizon thermality and quantum fluctuations in analog spacetimes.

Works Cited

- [1] R. M. Wald, *General Relativity*, University of Chicago Press, 1984.
- [2] G. C. Wick, “Properties of Bethe-Salpeter Wave Functions,” *Physical Review*, vol. 96, no. 4, pp. 1124–1134, 1954. doi:10.1103/PhysRev.96.1124
- [3] S. W. Hawking, “Particle Creation by Black Holes,” *Communications in Mathematical Physics*, vol. 43, no. 3, pp. 199–220, 1975. doi:10.1007/BF02345020
- [4] J. B. Hartle and S. W. Hawking, “Wave function of the Universe,” *Physical Review D*, vol. 28, no. 12, pp. 2960–2975, 1983. doi:10.1103/PhysRevD.28.2960
- [5] N. D. Birrell and P. C. W. Davies, *Quantum Fields in Curved Space*, Cambridge University Press, 1982. [6] M. Visser, *Acoustic Black Holes: Horizons, Ergospheres, and Hawking Radiation*, World Scientific, 1998.
- [7] C. Barceló, S. Liberati and M. Visser, “Analogue Gravity,” *Living Reviews in Relativity*, vol. 8, no. 1, p. 12, 2005. doi:10.12942/lrr-2005-12

Predicting Type 2 Diabetes Using Machine Learning: Insights from Feature Importance and Clinical Symptom Analysis By Luke Choi

Abstract

Diabetes is a chronic metabolic disorder affecting 10% of adults in the United States. Of those, 90% suffer from type 2 diabetes (Centers for Disease Control and Prevention, “Type 2 Diabetes”, 2024). Early and accurate prediction of diabetes is crucial for ensuring timely interventions and effective treatment strategies. This study leverages a publicly available dataset of approximately 500 diabetic patients to construct a predictive model of diabetes status using logistic regression. The predictor variables are a variety of clinical factors, such as polyuria, polydipsia, muscle stiffness. Emphasis is also placed upon model interpretation to identify key features associated with diabetes risk. Therefore, this model can be used to reliably predict whether a patient has diabetes, as well as enable a greater understanding of which conditions can be connected to this disease.

Introduction

Diabetes mellitus, a chronic metabolic condition characterized by impaired glucose regulation, remains a major health concern in the United States and worldwide. Current estimates suggest that more than 10% of U.S. adults live with diabetes, with many more at risk of developing the condition (Centers for Disease Control and Prevention, “Type 2 Diabetes”, 2024). Left untreated, diabetes can lead to severe complications including cardiovascular disease, neuropathy, kidney failure, and vision impairment (Centers for Disease Control and Prevention, “Put the Brakes on Diabetes Complications,” 2024). Consequently, early detection and accurate prediction of diabetes status are vital for guiding effective interventions and reducing long-term health burdens.

Statistical modeling offers a tool for identifying individuals at risk and uncovering factors most strongly associated with the disease. Machine learning models have been integrated into the variety of tools accessible to public health workers in past years, including those associated with disease diagnosis. In this study, we apply a type of statistical machine learning, logistic regression, to analyze a publicly available dataset of 520 newly diabetic patients to predict diabetes status as a function of various clinical conditions.

Logistic regression is a simple and transparent model that is highly interpretable, enabling clinicians and researchers to understand the relative contribution of different variables to the development of disease. The model provides easily quantifiable associations, such as the direction and magnitude of each predictor’s effect. The outputs can support evidence based decision-making by providing readable results. Additionally, the simplicity of logistic regression allows it to be compared against more complex approaches.

By combining predictive modeling with careful interpretation, this study aims to answer two key questions: (1) how effectively can survey and laboratory data predict diabetes status, and (2) which features show the strongest association with diabetes in this cohort? The results

demonstrate that logistic regression successfully achieves a predictive performance of 0.9645 compared to a randomized baseline of 0.5, and also identifies polyuria, polydipsia, and genital thrush as strong correlative factors.

Data Source

The *Early Stage Diabetes Risk Prediction* dataset (UCI Repository ID 529) was used, comprising 520 anonymous patients' self-reported conditions from newly diabetic or potentially diabetic individuals at Sylhet Diabetes Hospital, Bangladesh, uploaded in 2019.

These records include 17 variables (16 predictors + 1 outcome) (UCI Machine Learning Repository, 2019; PMC, 2014).

Key Variables

The dataset contains 17 variables. The target variable, 'Class', is a binary indicator of whether a patient tested positive or negative for type 2 diabetes. Of the predictor variables, 16 are binary clinical features, while 'Age' is the only continuous variable. These variables represent clinical characteristics associated with diabetes; Some of the included characteristics are polyuria– excessive urine production; polydipsia– excessive thirst; and muscle stiffness.

Data Preprocessing

The dataset was cleaned to identify missing values and ensure standardized formatting. No missing values were present, so no rows needed to be removed for inconsistency. The predictors consisted of 15 binary categorical variables and 1 continuous variable. Binary variables were encoded numerically, with 'No' assigned as 0 and 'Yes' as 1, and for sex, 'Female' was encoded as 0 and 'Male' as 1. The variable 'Age' was retained in its original numeric form and therefore not encoded.

After processing, there are 520 rows and 16 features, and one column for the patient's diabetes diagnosis. With respect to class balance, 320 (61.5%) of the patients had diabetes and 200 (38.5%) were healthy.

To train the model, 67% of the data was randomly assigned to the training set. The other 33% was dedicated for testing using sci kit learn's train test split function. The train-test split function is used to evaluate model performance by randomly partitioning the data into training and testing subsets. This approach allows the model to be trained on one portion of the data and then evaluated on a separate portion, providing an assessment of its predictive capability.

Model Selection: Logistic Regression

Logistic regression is a classification algorithm derived from linear regression that models the relationship between predictor variables and a binary outcome. It estimates the log-odds of the positive class (e.g., a patient having diabetes) as a linear combination of the input features. The linear output is then transformed using the logistic (sigmoid) function, which maps the result to a probability value between 0 and 1. This probability represents the model's

estimated likelihood that a given observation belongs to the positive class. Logistic regression is a helpful tool in diagnosis because it outputs the probability of a patient having diabetes, which provides enough information for medical specialists to make their own judgment regarding whether or not to move forward with an in-depth clinical test. The highly interpretable output allows for ease of access and generalized usage.

ROC (Receiver Operating Characteristic) curves are graphical plots utilized to evaluate the performance of diagnostic tests by plotting the true positive rate against the false positive rate. At each classification threshold for making a binary decision, the test calculates the true positive rate (sensitivity) and the false positive rate (1 - specificity) of the model's predictions, and then these pairs of values are plotted on the graph. Connecting the plotted points yields the ROC curve, and the area under the ROC curve (AUROC) provides an overall measure of the performance of the classifier.

An AUROC score of 0 represents a model that predicts every outcome incorrectly, meaning it has perfectly inverted predictions. Conversely, an AUROC score of 1 indicates a model with perfect discrimination between the positive and negative classes. An AUROC score of 0.5 is the score that a random predictor would achieve, and thus serves as a useful baseline for which to compare model performance.

Precision-recall curves visualize the relation between the recall (true-positive rate) and precision (positive predictive value) as the classification threshold changes. The precision is the ratio of the correctly predicted positives to the total number of predicted positives that the model made, while recall is the ratio of correctly predicted positive instances to all positive instances in the dataset. A model with high precision and high recall across thresholds produces a curve that approaches the top-right corner, indicating strong performance, especially in datasets with class imbalance. A random predictor would achieve a score equal to the prevalence of the positive class, which can again serve as a baseline.

Permutation Feature Importance analysis is a technique used to classify the extent to how much each input variable contributes to a model's predictive performance. It measures how much the model's accuracy changes when values of a single feature are randomly permuted, breaking the relationship between that feature and the target variable. If a permutation causes a large drop in performance, then that feature is important to the model and is given a higher importance value. Conversely, if there is little to no effect from the permutation, that feature is given a low value.

Software

All analysis was conducted with Python version 3.12.11 using Google Colab. Models were fit using scikit learn version 1.7.1.

Results

First, the performance of the model was evaluated using the AUPRC and AUROC metrics. As can be seen in Figure 2, the model achieved an area under the Precision-Recall

curve (AUPRC) of 0.9809, demonstrating high precision and recall performance. The high value reflects the model's ability to correctly identify positive cases (precision) and minimize false negatives (recall). It is substantially higher than the baseline prevalence of the positive class (0.615), indicating that the model is learning effectively relative to the baseline of a random classifier. In Figure 1, the ROC Curve (AUROC) is displayed. The AUROC is 0.9645, indicating a high capacity to correctly distinguish between diabetic and non-diabetic patients. Higher AUROC values indicate the model assigns consistently higher predicted probabilities to true positive cases comparatively to true negative cases. The model performance is again much higher than the baseline of 0.5, which is the performance of a random classifier.

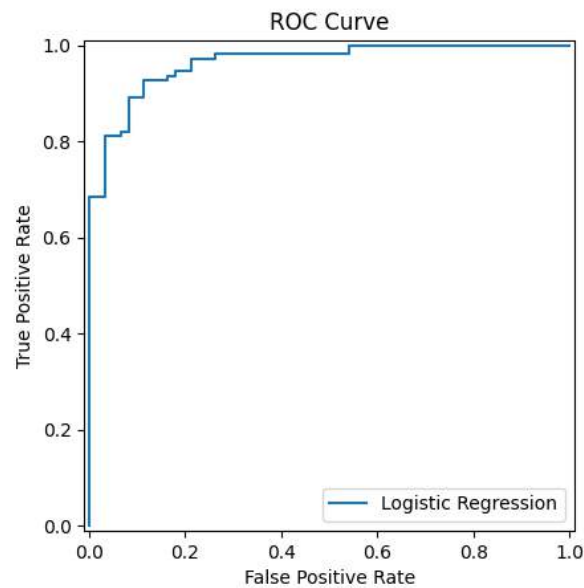


Fig 1: The ROC curve for the logistic regression model is plotted. The AUROC has a value of 0.965

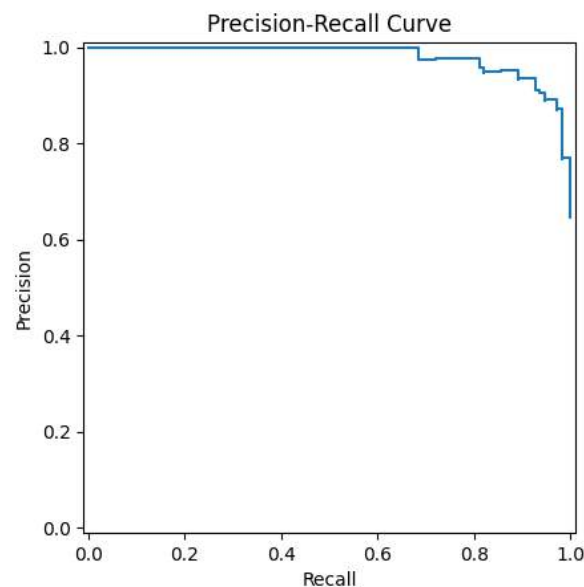


Fig 2: The Precision-Recall curve for the logistic regression model is plotted. The AUPRC has a value of 0.981

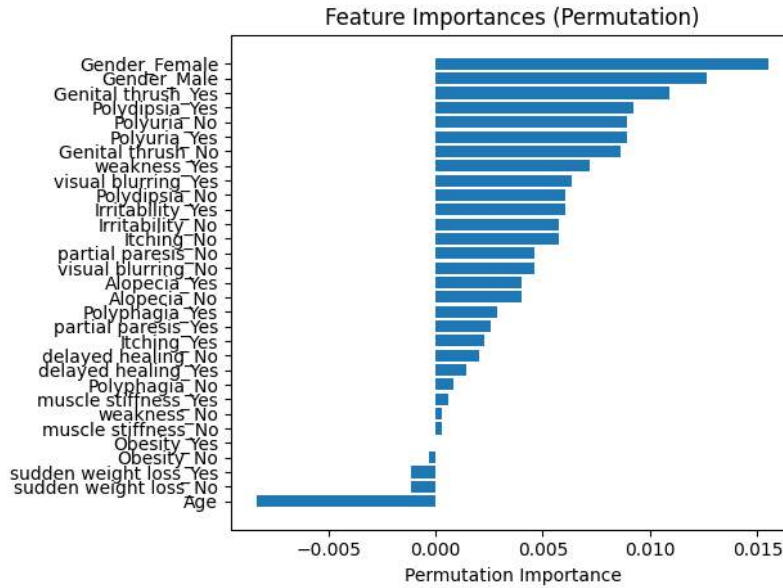


Fig 3: The Feature Importance curve is displayed above

As shown in Figure 3, the model demonstrated robust predictive performance. To interpret its behavior, a Permutation Feature Importance (PFI) test was conducted, revealing that the model relied most strongly on gender (0.0144), positive diagnoses of genital thrush (0.0083), polyuria (0.0077), and polydipsia (0.0074) in predicting type-2 diabetes,

Discussion

The model achieved a high diagnostic performance for type 2 diabetes, indicated by the AUPRC and the AUROC values. The high AUPRC also suggests that the model retains robust predictive accuracy when accounting for potential class imbalance (underrepresentation of diabetes positive or negative cases), ensuring positive diabetes cases are correctly identified with minimal false-positive classifications.

The PFI revealed that gender, polyuria, polydipsia, and genital thrush were the most strongly associated predictors of diabetes status. Gender-specific factors, such as differences in fat distribution and hormonal fluctuations according to the CDC, contribute to the distinct risk profiles observed in men and women with type 2 diabetes. Polyuria (frequent urination) and polydipsia (excessive thirst) are linked to hyperglycemia in patients as outlined in the *American Diabetes Association Standards of Care* (ADA, 2024). Furthermore, patients with diabetes are at increased risk for fungal infections, including genital thrush. Elevated blood glucose levels can lead to increased glucose excretion in urine and mucosal tissues, providing a nutrient-rich environment for *Candida* species to proliferate. High blood sugar levels can cause increased sweating and skin moisture, creating favorable conditions for yeast growth. In the context of type 2 diabetes, elevated blood glucose provides excess substrate (sugar) in tissues and secretions, promotes favourable local conditions (moisture, pH changes, mucosal glycogen), and combined with impaired immune function this creates an environment in which *Candida*

species (especially *C. albicans*, but also non-*albicans*) can proliferate more readily. The result is a higher risk of genital thrush (and other candidal infections) in individuals with diabetes, especially if glycaemic control is poor.

This study has several limitations. The small size of the dataset limits the generalizability of the results to a larger population, making it difficult to know how well the model will handle larger/more diverse datasets. Because the data was self-reported, it may also be subject to bias that might further skew the results given by the model in comparison to real values.

For future work, making use of larger, more diverse datasets will be beneficial for a model like this one, as it will have a wider range of subjects to train off of and improve its accuracy and reliability. In addition, alternative approaches could be explored that use different types of machine learning models. For instance, decision trees operate by starting at a root node and systematically following branches based on feature responses until reaching a final classification at a leaf node. When multiple decision trees are combined, they form a random forest model, which is widely used for its ability to reduce overfitting and improve predictive performance. Other ensemble methods, such as gradient boosting or XGBoost, could also be investigated to determine whether they offer further improvements in predictive accuracy and model interpretability. Furthermore, neural networks and deep learning models are increasingly becoming the de facto standard for analyzing complex and high-dimensional datasets. These models are capable of automatically learning intricate patterns and relationships within the data, which may enable even more accurate and robust predictions in future iterations of this research.

The findings from this study highlight the potential of machine learning models to serve as valuable tools in the early detection and risk assessment of type 2 diabetes. By identifying key predictive features such as polyuria, polydipsia, and genital thrush, this work could assist healthcare professionals in making more informed, data-supported clinical decisions. Integrating such models into clinical practice could help doctors identify at-risk patients earlier, prioritize diagnostic testing, and tailor interventions to individual risk profiles—ultimately improving patient outcomes and optimizing resource allocation in healthcare settings.

○

Works Cited

- Centers for Disease Control and Prevention. (2024, May 15). Type 2 diabetes. U.S. Department of Health and Human Services.
<https://www.cdc.gov/diabetes/about/about-type-2-diabetes.html>
- Centers for Disease Control and Prevention. (2024, May 15). Put the brakes on diabetes complications. U.S. Department of Health and Human Services.
<https://www.cdc.gov/diabetes/prevention-type-2/stop-diabetes-complications.html>
- Chang, A., Zhang, M., & Li, W. (2023). [Title of the article from PMC used for graph comparisons]. *PMC Articles*, 14, 1–15.
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10107388>
- UCI Machine Learning Repository. (2019). Early Stage Diabetes Risk Prediction dataset (Version 1) [Data set]. University of California, Irvine.
<https://archive.ics.uci.edu/ml/datasets/Early+Stage+Diabetes+Risk+Prediction>
- American Diabetes Association. (2024). Improving care and promoting health in populations: Standards of medical care in diabetes—2024. *Diabetes Care*, 47(Supplement 1), S11–S20.
https://diabetesjournals.org/care/article/47/Supplement_1/S11/153937/1-Improving-Care-and-Promoting-Health-in
- Healthline. (2024). Diabetes and yeast infections: How diabetes affects risk.
<https://www.healthline.com/health/diabetes/diabetes-and-yeast-infections#outlook>
- Centers for Disease Control and Prevention. (2023). Diabetes and men. U.S. Department of Health & Human Services.
<https://www.cdc.gov/diabetes/risk-factors/diabetes-and-men.html>
- Banner Health. (2024, March 5). What to know about yeast infections if you have diabetes. Banner Health.
<https://www.bannerhealth.com/healthcareblog/teach-me/what-to-know-about-yeast-infections-if-you-have-diabetes>
- Centers for Disease Control and Prevention (CDC). (2021, July 22). Sexually transmitted infections treatment guidelines, 2021: Vulvovaginal candidiasis. CDC.
<https://www.cdc.gov/std/treatment-guidelines/candidiasis.htm>
- Darwazeh, A. M. G., & Al-Doughmi, I. A. (2021). Candida infections in patients with diabetes mellitus. *Journal of Diabetes Research*, 2021, Article 8893836.
<https://doi.org/10.1155/2021/8893836>
- De Bernardis, F., Graziani, S., Tirelli, F., & Antonopoulou, S. (2021). Candida infections and the diabetic foot: How strong is the association? *Pathogens*, 10(2), 164.
<https://pmc.ncbi.nlm.nih.gov/articles/PMC7980094/>
- Patil, S., Rao, R. S., Majumdar, B., & Anil, S. (2018). Clinical appearance of oral Candida infection and therapeutic strategies. *Frontiers in Microbiology*, 9, 881.
<https://pmc.ncbi.nlm.nih.gov/articles/PMC6352194/>

Appendix

Variables used:

- Class: Type II diabetes diagnosis
- Age: numeric (1–65 years)
- Gender: categorical (Male/Female)
- Clinical measurements (binary, Yes/No), including:
 - Polyuria (excess urination)
 - Polydipsia (excess thirst)
 - Sudden weight loss
 - Weakness
 - Polyphagia (excess hunger)

 - Genital thrush
 - Visual blurring
 - Itching
 - Irritability
 - Delayed healing
 - Partial paresis
 - Muscle stiffness
 - Alopecia
 - Obesity

Microloans, Fintech, and Female Entrepreneurship: A Conceptual Study of Women-led MSMEs in Kazakhstan By Maren Webb

Abstract

This conceptual paper examines how gender norms, institutional environments, and geography intersect to shape the development and performance of women-led micro, small, and medium-sized enterprises (MSMEs) in Kazakhstan, with specific focus on rural contexts. Drawing on Eagly's Social Role Theory, Becker's Human Capital Theory, and North's New Institutional Economics, the paper develops a dual-level framework that distinguishes between macro-and micro-level factors operating across rural and urban settings. It argues that while microfinance and fintech can reduce structural barriers, such as limited access to capital and geographic isolation, their effectiveness depends critically on local institutional trust, digital literacy, and gender-role expectations. Four propositions are advanced to explain these dynamics: (1) restrictive gender norms amplify weak institutional trust in rural regions; (2) urban ecosystems enhance women's agency through greater access to resources and networks; (3) fintech mitigates spatial barriers but fails to overcome cultural ones; and (4) the impact of financial inclusion initiatives is contingent upon institutional credibility and digital competence. By situating Kazakhstan within global debates on gender and financial inclusion, this paper contributes a regionally grounded framework for understanding uneven entrepreneurial outcomes. Finally, it offers policy insights for integrating digital finance, capacity-building, and gender-sensitive institutional reform to foster inclusive, sustainable growth among women entrepreneurs.

Introduction

Micro, small, and medium-sized enterprises (MSMEs) are a cornerstone of Kazakhstan's economic diversification strategy, contributing over one-third of the national GDP and employing more than 4 million people (World Bank). The sector is geographically concentrated, with urban hubs such as Almaty, Astana, Shymkent, Karaganda, and Aktobe hosting the majority of MSMEs, primarily operating often in trade, services, construction, and manufacturing. These enterprises benefit from stronger infrastructure, better access to finance, and greater integration into national and international value chains (Damu).

In rural areas, including the Almatinskaya, Turkistan, Kostanay, East Kazakhstan, and Zhambyl regions, MSMEs are typically smaller, more informal, and concentrated in agriculture, agro-processing, local retail, and handicrafts (World Bank). While they are less connected to formal supply chains, they play an essential role in sustaining local economies, diversifying income beyond subsistence farming, and absorbing seasonal labor. However, rural enterprises face structural barriers such as underdeveloped infrastructure, limited market access, and lower digital penetration, which collectively restrict growth potential (UNDP Kazakhstan).

The rural-urban divide in MSME performance reflects not only geographic and infrastructural disparities but also institutional, cultural, and social contexts. Urban MSMEs

benefit from more developed entrepreneurial ecosystems—including incubators, training programs, and financial institutions—while rural enterprises often rely on informal financing and social capital to sustain operations. Programs like the Damu Entrepreneurship Development Fund have targeted both rural and urban sectors, but uptake varies due to differences in administrative capacity, institutional trust, and physical and digital accessibility (Damu).

Conceptual Questions and Contribution

This paper asks: How do gender norms and institutional environments interact to shape women entrepreneurs' ability to leverage microfinance and fintech across rural and urban contexts in Kazakhstan?

The paper's contribution is twofold. First, it extends entrepreneurship research into the underexplored Central Asian context, situating Kazakhstan within broader global debates on gender, finance, and digital innovation. Second, it develops a multi-level conceptual model integrating gender theory, human capital, and institutional economics to explain uneven adoption and outcomes of financial tools among women entrepreneurs. This model highlights how contextual boundary conditions—cultural norms, institutional trust, and digital literacy—shape the effectiveness of financial inclusion mechanisms (North).

Theoretical Framework

This study adopts a dual-theoretical approach that integrates Eagly's Social Role Theory and Becker's Human Capital Theory, supplemented by insights from North's New Institutional Economics.

Eagly's Social Role Theory (1987) explains how entrenched gender norms and role expectations influence women's entrepreneurial participation. In Kazakhstan, these roles are often more rigidly prescribed in rural areas, constraining risk-taking and access to entrepreneurial networks. Liñan et al. (2020) extend this logic by demonstrating how gender-role orientation (GRO) interacts with perceived regional culture (PRC) to shape women's business decisions.

Becker's Human Capital Theory (1994) emphasizes that investments in education, skills, and health increase productivity and economic output. However, when women face barriers to accessing finance, training, or technology, their human capital remains underdeveloped and underutilized, leading to persistent economic inefficiencies (Becker).

North's New Institutional Economics (1992) emphasizes that institutional quality—including formal rules, informal norms, and enforcement mechanisms—shapes market participation and opportunity structures. This perspective is particularly relevant to understanding why and how microfinance and fintech adoption diverge between rural and urban contexts in Kazakhstan.

Together, these theoretical perspectives elucidate the “compound barriers” facing women entrepreneurs in rural settings and illustrate the conditions under which microfinance and fintech may mitigate—or reinforce—these constraints (Eagly and Wood; North).

Conceptual Framework

Building on these theoretical foundations, this paper advances an intersectional conceptual framework that examines the overlapping influences of gender and geography on the entrepreneurial experiences of women-led MSMEs in Kazakhstan. The model employs a two-by-two matrix distinguishing macro-level structural conditions from micro-level lived experiences across urban and rural contexts. This structure enables a nuanced understanding of how institutional, cultural, and technological factors interact to shape women's entrepreneurial outcomes.

Proposition 1 (P1): In rural contexts, restrictive gender norms amplify the negative effects of weak institutions, limiting women's entrepreneurial participation even when resources are available (Eagly and Wood; North).

Proposition 2 (P2): Urban contexts, characterized by stronger entrepreneurial ecosystems, enable fintech and microfinance to enhance women's business agency and growth, conditional upon their accumulated human capital (Becker).

Proposition 3 (P3): Fintech adoption reduces structural barriers such as distance and transaction costs, but persistent cultural constraints—including norms and trust—remain binding unless explicitly addressed (Becker).

Proposition 4 (P4): The positive impact of microfinance and fintech is contingent on institutional trust and digital literacy, both of which vary significantly between rural and urban women. (North).

Figure 1. *2x2 Conceptual Framework*

Conceptual Framework: Differentiates between macro-level structural conditions and micro-level lived experiences		
	Macro	Micro
Urban	National and municipal level policies, state-led MSME support programs, and market access in major cities such as Almaty, Astana, Shymkent, and Karaganda (P4)	Professional women entrepreneurs navigating gendered expectations within competitive, male-dominated sectors while leveraging urban networks and digital tools (P2)
Rural	Regional infrastructure quality, investment patterns (P3, P4)	Women in remote, rural regions balancing household responsibilities with business management, facing challenges in loan acquisition, digital connectivity, and access to formal markets (P1)

Note: Author's conceptual model, informed by Eagly and Wood (1987), Becker (1994), and North (1992).

This framework emphasizes the intersection of gender and geography, highlighting how rural women face compound constraints. Geographical isolation limits infrastructure development and market opportunities, while entrenched gender norms restrict entrepreneurial decision making and risk taking. Although urban women entrepreneurs also encounter gender bias, they often exercise greater agency because of stronger access to resources, networks, and institutional support.

Table 1. *Country-Level Framework of Gender, Geography, and Institutional Conditions in Kazakhstan*

	Population	Geography	Institutional Conditions
Almaty (City)	• Total Population: 2,228,675 • 100% Urban Population	• Distance to major city center: N/A	• Access to traditional banks: High access; of Kazakhstan's 38 banks containing 368 branches, nearly 300 are located in the 14 largest cities • Ease of starting a business rank: 3
Astana (City)	• Total Population: 1,430,105 • 100% Urban Population	• Distance to major city center: N/A	• Access to traditional banks: Same as Almaty • Ease of starting a business rank: 1
Kyzylorda	• Total Population: 846,104 • 52.8% Rural Population	• Distance to closest major city center (Shymkent): 456 km	• Access to traditional banks: Low/medium access; of Kazakhstan's 368 bank branches, only 72 branches are accessible in rural area • Ease of starting a business rank: 2
Mangistau	• Total Population: 805,156 • 53.2% Rural Population	• Distance to closest major city center (Shymkent): 2,212 km	• Access to traditional banks: Same as Kyzylorda • Ease of starting a business rank: 4
Zhetisu	• Total Population: 694,325 • 55.6% Rural Population	• Distance to closest major city (Almaty): 270 km	• Access to traditional banks: Same as Kyzylorda • Ease of starting a business rank: T-5
Zhambyl	• Total Population: • 56.6% Rural Population	• Distance to nearest major city (Shymkent): 182 km	• Access to traditional banks: Same as Kyzylorda • Ease of starting a business rank: 6

	Population	Geography	Institutional Conditions
Turkistan	• Total Population: 2,154,041 • 75.2% Rural Population	• Distance to nearest major city (Shymkent): 186 km	• Access to traditional banks: Same as Kyzylorda • Ease of starting a business rank: Not ranked in Doing Business Report
Almaty (Region)	• Total Population: 1,560,124 • 80.4% Rural	• Distance to nearest major city (Almaty): 100 km	• Access to traditional banks: Same as Kyzylorda • Ease of starting a business rank: T-4

Note: Data compiled by author from multiple sources, including the World Bank and Kazakhstan National Bureau of Statistics

The information presented in Table 1 reinforces my conceptual framework by illustrating systematic differences between urban and rural settings for women entrepreneurs. Urban centers such as Almaty and Astana, with fully urban populations, dense banking networks, and higher ease-of-starting-business rankings, reflect more developed infrastructure and institutional capacity. In contrast, regions such as Turkistan, Zhetisu, and Almaty (Region) have predominantly rural populations, greater physical distance from major cities, and fewer accessible bank branches. These patterns underscore the structural and institutional limitations confronting entrepreneurs in rural areas. The data further demonstrate that institutional accessibility plays a central role in shaping financial inclusion: even when fintech or microfinance tools are available, their uptake is often constrained by geographic isolation, limited infrastructure, and minimal exposure to formal financial systems.

The theoretical framework and conceptual model collectively give rise to four propositions, which are outlined in the section below.

Proposition Development

Women's Entrepreneurship in Kazakhstan (Rural vs. Urban)

Women own approximately 42 percent of MSMEs in Kazakhstan, but their entrepreneurial activity is shaped by complex intersections of gender norms, regional culture, and access to resources (UNDP Kazakhstan). Urban women entrepreneurs, particularly those in Almaty, Astana, and Shymkent, operate in more competitive and diversified markets, benefiting from greater exposure to business networks, training opportunities, and formal financial institutions (EBRD). In contrast, rural women, especially in Turkistan, Pavlodar, and East Kazakhstan, are often concentrated in lower-margin, locally oriented enterprises in agriculture, handicrafts, and small-scale retail (World Bank).

As previously mentioned, Eagly's Social Role Theory (1987) explains how entrenched gender norms and role expectations shape women's entrepreneurial participation. In Kazakhstan,

these roles are especially restrictive in rural regions, constraining risk-taking and access to entrepreneurial networks. Liñan et al. (2020) extend this logic by demonstrating that GRO interacts with PRC to influence women's business decisions. The GRO framework highlights the importance of psychological identity in women's entrepreneurial engagement (Liñan et al.). Women whose self-concepts align with traditionally "masculine" or androgynous traits, such as assertiveness and risk-taking are more likely to pursue entrepreneurship, particularly in supportive PRCs. However, rural Kazakhstan often exhibits conservative PRCs that reinforce traditional gender roles, limiting women's engagement even when financial tools are available (Liñan et al.). Feminine GROs, emphasizing communal and regional traits, are less responsive to structural reforms, suggesting that institutional changes alone may not close participation gaps.

These dynamics indicate that interventions promoting women's entrepreneurship in rural Kazakhstan must address both structural barriers and internalized gender norms, by combining financial access with targeted programs that build confidence, agency, and leadership capacity (UNDP Kazakhstan). Thus, I advance the following propositions:

P1: In rural contexts, restrictive gender norms amplify the negative effects of weak institutions, limiting women's entrepreneurial participation even when resources are available (Eagly and Wood; North).

P2: Urban contexts, characterized by stronger entrepreneurial ecosystems, allow fintech and microfinance to enhance women's business agency and growth, conditional on human capital (Becker).

Women Entrepreneurs, Microfinance, and Fintech in Kazakhstan (Rural vs. Urban)

Microfinance and fintech represent critical pathways for improving women entrepreneurs' access to capital in Kazakhstan, particularly in rural areas where traditional banking infrastructure remains limited (Asian Development Bank). Microfinance in Kazakhstan frequently operates as a hybrid system, combining formal lending with informal practices, thereby supporting adaptive entrepreneurial behavior in contexts of institutional mistrust. Borrowers often prioritize repayment reliability over high-risk investments, using loans primarily to stabilize income rather than to aggressively expand operations.

Urban women entrepreneurs increasingly adopt fintech tools such as mobile payments, e-commerce platforms, and digital microloans, to reach broader markets and streamline operations. In rural contexts, however, fintech's potential is constrained by low digital literacy, uneven internet connectivity, and persistent gender norms that discourage women from engaging with unfamiliar technologies. North's (1992) New Institutional Economics framework helps explain this divergence: institutional quality, both formal (banking regulations, property rights) and informal (cultural norms, trust networks), determines how technology and finance are adopted.

While fintech can mitigate institutional constraints by reducing transaction costs and circumventing physical barriers, it can also exacerbate existing inequalities if digital access and literacy gaps are not addressed. Consequently, tailored capacity-building, local outreach, and

trust-building initiatives are essential for realizing fintech's potential among rural women-led MSMEs. Thus, I propose:

P3: Fintech adoption reduces structural barriers such as distance and transaction costs, but cultural barriers related to norms and trust remain binding unless explicitly addressed (Becker).

P4: The positive impact of microfinance and fintech is contingent upon institutional trust and digital literacy, both of which vary significantly between rural and urban women entrepreneurs (North).

Discussion

The integrated two-by-two framework illuminates how women's entrepreneurship in Kazakhstan cannot be explained solely by structural conditions or individual agency. Rather, the intersection of gender norms, institutional quality, and geography determines who participates in the modern economy and how effectively they do so. The framework underscores that empowerment is a multi-level process: micro-level factors such as skills and confidence must be coupled with macro-level reforms that expand access, trust, and opportunity. It is also important not to conflate urban and rural experiences. The challenges faced by urban entrepreneurs differ substantially from those encountered by rural entrepreneurs; therefore, analyses of women's participation must account for the contextual environments in which they operate. The synthesis of macro/micro and urban/rural factors advances theory by linking Social Role Theory, Human Capital Theory, and New Institutional Economics within a single framework that captures the nuanced experiences of women entrepreneurs. This framework creates a dynamic depiction of how women's agency, institutional structures, and technology evolve together across Kazakhstan's diverse regional landscapes. Building on this foundation, the following discussion interprets each quadrant of the conceptual model in relation to the propositions (P1-P4).

At the micro-rural level, women entrepreneurs in Kazakhstan face overlapping social and structural barriers that limit their ability to access and effectively utilize financial tools. Rural women often balance household responsibilities with business management, leaving limited time for capacity building or networking (Mashapure et al.). Consistent with my conceptual model and Proposition 1 (P1), restrictive gender norms are likely to magnify the negative effects of weak institutions, creating "compound barriers" that restrict participation even when formal resources, like microloans or financial services, are available. Secondly, low levels of digital and financial literacy, coupled with limited exposure to entrepreneurial role models, further erode confidence in using microfinance and fintech platforms. Based upon the integration of Eagly's Social Role Theory and my conceptual framework, it is illuminated how cultural expectations surrounding women's domestic and caregiving roles constrain their perceived suitability for entrepreneurship, particularly in regions where institutional and infrastructural support is limited.

In contrast, the micro-urban context offers women entrepreneurs more autonomy and agency, aligning with my Proposition 2 (P2): stronger ecosystems enable fintech and microfinance to enhance women's business growth, conditional on human capital. Women in

Almaty, Astana, and Shymkent often benefit from access to professional networks, higher education, and skills training that strengthen technological competence and self-efficacy. Becker's Human Capital Theory suggests that these women's prior investments in education and work experience increase their capacity to use digital financial tools and credit systems effectively. Nevertheless, even in urban centers, gendered expectations continue to shape entrepreneurial behavior and experiences, revealing that social constraints persist, though they are mitigated by resource access and institutional inclusion.

At the macro-rural level, structural barriers, including underdeveloped infrastructure, weak digital connectivity, and sparse banking networks, amplify micro-level disadvantages. Thus, as articulated in Propositions 3 and 4 (P3-P4), these barriers demonstrate that while fintech can theoretically reduce spatial and transactional limitations, its potential is constrained where institutional trust and digital literacy are low. Rural women often express skepticism towards digital finance technologies, due to inconsistent regulations and past negative lending experiences. From the North's New Institutional Economics, the absence of credible, trustworthy institutions impedes the diffusion of innovative financial systems. In these settings, cultural barriers, not merely technological or logistic ones, remain the most formidable obstacles to inhibit inclusion.

Conversely, macro-urban level environments demonstrate how robust institutional and policy frameworks can support women's entrepreneurial participation. Urban regions benefit from national programs such as the DAMU Entrepreneurship Fund and the EBRD Women in Business initiatives, which offer credit guarantees, training, and digital literacy programs. These state and market-driven interventions strengthen confidence in financial systems and expand women's access to digital microfinance tools, confirming the patterns outlined in Proposition 4 (P4). The macro-urban quadrant therefore represents the convergence of institutional support and individual agency, showing that when infrastructural and institutional systems align, women entrepreneurs can fully leverage fintech and microfinance to scale their business to their maximum potential. Taken together, these findings emphasize that successful inclusion strategies must be both structurally enabling and individually empowering.

Practical Implications

There are four primary practical implications of my research. First, partnerships among MFIs, NGOs, and local women's associations can serve as intermediaries, translating formal financial mechanisms into accessible and trusted practices for women who operate largely within informal or household-based economies. Programs introducing microfinance and fintech tools should incorporate capacity-building components, including digital literacy training, peer mentoring, and culturally sensitive outreach. Second, policymakers should implement trust-building measures to strengthen the credibility of financial institutions in rural regions. Transparent communication about loan terms, grievance mechanisms, and data protection can reduce skepticism that deters women from engaging with formal lenders and digital platforms. Embedding these initiatives within community-based networks, rather than imposing them

through top-down bureaucratic channels, is likely to increase legitimacy and participation. Third, at the macro level, government and development agencies should prioritize expanding financial and digital infrastructure in underserved areas. This includes extending broadband coverage, improving mobile payment interoperability, and incentivizing banks and fintech providers to operate in rural districts. Importantly, institutional expansion must be paired with gender-responsive program design, recognizing that rural women face distinct barriers. Financial inclusion initiatives that neglect these contextual realities risk reinforcing inequality rather than reducing it. Fourth, in urban centers, where fintech ecosystems are already emerging, the policy priority shifts toward consolidating and scaling inclusion gains. This may involve developing mentorship networks, women-focused venture accelerators, and targeted investment programs supporting female-led digital enterprises. Collaboration among fintech firms, universities, and municipal governments could further strengthen women's participation not only as consumers but as innovators and leaders within the digital economy.

Future research should adopt my two-by-two framework to explore how gender norms, financial technologies, and institutional contexts intersect to shape women's entrepreneurial experiences. Comparative studies within Central Asian could reveal how varying policy environments and cultural factors influence entrepreneurial outcomes. Moreover, mixed-methods research designs could capture how fintech adoption, microloan access, and capacity-building initiatives affect women-led MSMEs over time. Expanding both the empirical and theoretical scope of research will enhance understanding of how to foster inclusive financial ecosystems that will genuinely empower women in rural and transitional economies.

Ultimately, fintech holds the transformative potential to bridge spatial and institutional divides, but only when micro-level barriers are addressed alongside macro-level reforms. Policies that integrate these dimensions can transform financial inclusion from a formal policy goal into a lived reality for women across Kazakhstan's urban and rural landscapes. Such an approach not only advances gender equity but also furthers the country's broader agenda of inclusive, innovation-driven economic growth.

Conclusion

This study has argued that women's entrepreneurship both in city and rural areas in Kazakhstan is shaped by the interaction of gender norms, institutional quality, and geography. By integrating Social Role Theory, Human Capital Theory, and New Institutional Economics within a two-by-two macro/micro and urban/rural framework, it demonstrates that structural reform and individual empowerment must advance together for financial inclusion to be meaningful. While fintech and microfinance can expand access to capital, their success ultimately depends on trust, literacy, and cultural adaptation. The conceptual model proposed here moves beyond the urban-rural dichotomy to show that women's economic participation is both a product of and a catalyst for institutional transformation. Sustained progress will require policies that not only connect women to financial systems but also embed equality, capability, and confidence at the heart of Kazakhstan's evolving entrepreneurial landscape.

Works Cited

- Annual Report on SMEs Development in Kazakhstan and Its Regions. DAMU, 2019, damu.kz/upload/iblock/192/Damu_Book_2019_EN.pdf. Accessed 13 Aug. 2025.
- Beck, Thorsten. *Fintech and Financial Inclusion: Opportunities and Pitfalls*. ADBI Working Paper 1165, Asian Development Bank Institute, 2020. Asian Development Bank, <https://www.adb.org/publications/fintech-financial-inclusion-opportunities-pitfalls>
- Becker, Gary S. "Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education." National Bureau of Economic Research, 1975, www.nber.org/system/files/chapters/c3730/c3730.pdf. Accessed 21 Oct. 2025.
- Eagly, Alice H., and Wendy Wood. "Social Role Theory." *Handbook of Theories of Social Psychology*, edited by Paul A. M. Van Lange, Arie W. Kruglanski, and E. Tory Higgins, vol. 2, Sage Publications, 2012, pp. 458–476. <https://doi.org/10.4135/9781446249222.n49>
- Kazakhstan Demographic Statistics. Kazakhstan Bureau of National Statistics, stat.gov.kz/en/industries/social-statistics/demography/publications/40410/. Accessed 21 Oct. 2025.
- "A Lifeline for Small Businesses in Kazakhstan to Survive and Thrive." Asian Development Bank, 8 Nov. 2024, www.adb.org/results/lifeline-small-businesses-kazakhstan-survive-thrive. Accessed 21 Oct. 2025.
- Liñán, Francisco, Iván Jaén, and Daniel Martín. "Does Entrepreneurship Fit Her? Women Entrepreneurs, Gender-Role Orientation, and Entrepreneurial Culture." *Small Business Economics*, vol. 58, no. 2, 2022, pp. 1051–1071. Springer, <https://doi.org/10.1007/s11187-020-00433-w>.
- Mashapure, R., et al. "Challenges Hindering Women Entrepreneurship Sustainability in Rural Livelihoods: Case of Manicaland Province." *Cogent Social Sciences*, vol. 8, no. 1, 2022, <https://doi.org/10.1080/23311886.2022.2132675>
- North, Douglass C. "The New Institutional Economics and Development." *Econ WPA Archive*, econwpa.ub.uni-muenchen.de/econ-wp/eh/papers/9309/9309002.pdf. Accessed 21 Oct. 2025.
- "Rural Women's Businesses – First Steps to Coming of Age." UNDP Kazakhstan, 13 Oct. 2022, www.undp.org/kazakhstan/stories/rural-womens-businesses-first-steps-coming-age. Accessed 22 Oct. 2025.
- "Starting a Business - Kazakhstan." World Bank Subnational Studies, subnational.doingbusiness.org/en/data/exploretopics/starting-a-business/kazakhstan. Accessed 21 Oct. 2025.
- Usov, Anton. "EBRD Supports Women Entrepreneurs in Kazakhstan." EBRD, www.ebrd.com/home/news-and-events/news/2025/ebrd-supports-women-entrepreneurs-in-kazakhstan.html#. Accessed 22 Oct. 2025.
- The World Bank Group. "The Role of Postal Networks in Expanding Access to Financial Services." *World Bank Documents and Reports*, 2004, documents1.worldbank.org/curated/en/306051468153556802/pdf/694490ESW0P0850PUBLIC0000Kazakhstan.pdf. Accessed 21 Oct. 2025.

The Effects of Neurologic Music Therapy Techniques on Patients with Cerebral Palsy

By Zachary W. Ong

Abstract

Cerebral palsy (CP) is a nonprogressive neuromotor disorder that negatively affects the quality of life of those affected by it. Neurologic music therapy (NMT) – specifically rhythmic auditory stimulation (RAS), therapeutic instrumental music performance (TIMP), and patterned sensory enhancement (PSE) – has shown potential as an adjunct to conventional CP treatments. This review article synthesized peer-reviewed studies published between 2007 and 2021. A total of six studies were identified through searches of Google Scholar, PubMed, and other journal archives. The studies indicate that RAS in conjunction with physical therapy is effective at improving gait parameters such as stride length, velocity, symmetry, overall balance, and, in some cases, improved psychological well-being. TIMP demonstrated notable improvements in the motor function of upper extremities and improvements in grip strength in adolescents, although results in adults were mixed. PSE treatments in conjunction with sit-to-stand exercises yielded improvements in motor function, functional skills, knee extensor power, and decreases in jerkiness. Study limitations are small sample sizes, population samples that lack diversity, short intervention periods, and differing methodologies, which limit comparisons between studies. Overall, current evidence suggests that NMT can improve physical and psychological outcomes in individuals with CP as a complement to conventional treatment. However, future studies with larger sample sizes, more diverse sample populations, longer interventions, and NMT alone are necessary to investigate the effects of NMT to its full extent.

Cerebral Palsy

Cerebral palsy (CP) is a nonprogressive neuromotor disorder that is caused by injury to the developing brain (Patel et al.). Injuries occur around 92% of the time in the perinatal period (Vitrikas et al.) though abnormal development and damage to the brain during the prenatal, perinatal, and postnatal period can cause CP (Hallman-Cooper and Rocha Cabrero). CP occurs in around two to three out of 1,000 live births (Jan) and is the most common childhood disability (Hallman-Cooper and Rocha Cabrero). It can cause disabilities such as mental retardation, epilepsy, spasticity, feeding difficulties, vision and hearing impairment, and difficulties with movement (Jan). The result of these disabilities is reduced physical activity and a lowered quality of life (Graham et al.).

Since movement disorders are the most common disorders, we will primarily be discussing studies with individuals with motor disorders. The movement disorders related to CP are athetoid (uncontrolled and jerky movements), ataxic (impaired coordination), hypotonic (lack of muscle tone), and spastic (muscle stiffness and increased muscle tone) (Shamsoddini et al.; Tugui and Antonescu). This review will primarily focus on spasticity as it is the most common motor disability among children with CP (Shamsoddini et al.) with an occurrence rate of 80% among children with CP (Vitrikas et al.). Spastic CP can be further classified by the body area

affected as hemiplegia, diplegia, tetraplegia, and quadriplegia (Vitrikas et al.). Management of spasticity is imperative as spasticity can negatively affect daily life and be a safety concern as it causes abnormal gait, muscle pain or spasms, and inhibition of muscle growth (Shamsoddini et al.). Individuals with CP also experience various psychological problems such as negative emotional image and low self-esteem (Efraimidou et al.). There are many ways to manage the effects of CP such as neurodevelopmental treatments, physical therapy, muscle relaxants, surgery, and external aids (Krigger). However, these treatments are solely implemented to minimize the negative effects of the disorders, not a specific treatment for the brain insults causing the motor problems (Papavasiliou).

Music Therapy

Music therapy is the clinical and evidence-based use of music to promote wellness, manage stress, alleviate pain, improve communication, and promote physical rehabilitation (American Music Therapy Association). Music therapy utilizes many different techniques that include the use of rhythmic, harmonic, and melodic sounds (Vinolo-Gil et al.). Currently, music therapy is primarily used in mental health settings such as schools, geriatric facilities, and private practice (Kern and Tague) but it shows promise in the management of cerebral palsy by itself or in conjunction with physical therapy (Vinolo-Gil et al.). Music therapy includes many different therapeutic techniques, a subset which is called Neurologic Music Therapy (NMT). NMT includes techniques such as rhythmic auditory stimulation (RAS), patterned sensory enhancement (PSE), and therapeutic instrumental music performance (TIMP) (Thaut and Hömberg).

Rhythmic Auditory Stimulation

RAS is the NMT technique that uses the rhythmic nature of music to facilitate rehabilitation, development, and maintenance of movements that are biologically rhythmical such as gait. RAS is built on four neurological principles: rhythmic entrainment, auditory priming, cueing of the movement period, and stepwise limit-cycle entrainment. Rhythmic entrainment is the ability for the motor system to become synchronized with incoming sensory information from the auditory nervous system. Auditory priming is the ability of anticipatory auditory cues to improve the synchronization and stimulation of the systems related to rhythmic entertainment. Cueing of the movement period is the use of a continuous and consistent rhythm to improve synchronization in auditory-motor entrainment. Stepwise limit-cycle entrainment is the process of entraining an abnormal step cadence and then normalizing the cadence through a stepwise progression protocol (Thaut and Rice).

2007 Study by Eunmi Kwak

In 2007, Kwak investigated the effects of RAS on the gait of children with spastic CP in the Chungju Hae-Hwa School for children. This school, based in Korea, specialized in students with physical disabilities. A total of 30 participants, aged 6 to 20 years old, participated in a

controlled experiment involving 3 different groups. The groups included a control group who only received physical therapy gait training from a physical therapist, a therapist-guided training group (TGT) who received both physical and RAS therapy from a physical and a music therapist, and a self-guided group (SGT) who received both physical and musical therapy from a physical therapist and followed a self-guided RAS training program. Gait training was held in unrestricted spaces in the school such as playgrounds, hallways, and slopes (not stairs). The researchers used a MIDI program to provide tempo changes to recorded music to accommodate the participants' varied cadences. They used three basic melodies that all had a uniform 4/4 beat pattern and a tempo ranging from 80 bpm to 120 bpm. A metronome was used to confirm the tempos during the experiment. Drums and clapping were used to emphasize the beat in the music. Different acoustics from the drum were used to indicate the use of the right foot and the left foot. To get accurate measurements the researchers used a stride analyzer that records foot-floor contact data and then calculates and compares gait parameters. With the stride analyzers attached to the participants, participants were asked to walk at their natural tempo for pre-tests and post-tests along a measured walkway. The researchers then used the pre-tests to determine the tempo in RAS training. The tempo of music used in RAS training was 5% higher than the participants' pretest walking tempo during the first week. The tempos were then increased by 10% the second week and by 15% in the third week.

Results

Significance in the findings varied. The stride length was shown to be statistically significant in the TGT group with an increase of 29.48%. The velocity was shown to be statistically significant in the TGT group with an increase of 36.49%. The symmetry was shown to be statistically significant in the TGT group with an improvement of 16.97%. Results in the SGT group were shown to be statistically insignificant. Additionally, the change in cadence was shown to be statistically insignificant between pre and post-test for all groups. These results show that the use of physical therapy and RAS in conjunction along with the presence of both a physical and a music therapist can improve the gait of children with CP through their stride length, velocity, and symmetry (Kwak).

2016 Study by Vasiliki Efraimidou et al.

In 2016, Efraimidou et al. investigated the effects of a RAS and movement in combined program on adults with spastic CP at a sports club in Thessaloniki. A total of 10 participants, males aged 22 to 51 years old and in throwing events were randomly divided into 2 groups of 5. The first group was the intervention group (IG) who followed an 8-week RAS and movement program that occurred 2 times per week for 50 minutes a session. The second group was the control group (CG) who followed a regular ball and disc training program. The participants had their gait, balance, and psychological parameters measured before and after the program was implemented. The researchers used two tests to measure gait: the Timed Up and Go test (TUG) which measured gait time and the 10 Meter Walk Test (10 MWT) which measured the normal

and fast gait speed. To measure the balance of the participants before and after the researchers used The Berg Balance Scale (BBS) and the EPS pressure platform. To assess the psychological parameters of the participants the Self Esteem Scale (RSE) questionnaire was used along with the Profile of Mood States (POMS). In the treatment sessions music at a tempo of 70 bpm to 90 bpm with a 4/4 music meter was played on a sound CD machine. Participants performed physical movements in synchrony with the music with the goal of integrating the rhythm of the music into these movements.

Results

The results showed improvements in many aspects of a patient's gait and psychological wellbeing in the IG. Gait time in the TUG test was statistically significant with improvements from around 10.2 seconds to around 7.5 seconds. In the 10 MWT statistical significance was shown in normal gait speed with improvements from 0.59 m/s to 0.73 m/s and fast gait speed with improvements from 0.86 m/s to 1.05 m/s. In the BBS statistical significance was seen in the balance score which improved from around 41 to around 49.8. In the RSE statistical significance was seen in the self-esteem score which improved from 30.4 to 36.4. Finally, in the POMS statistical significance was shown in intention (5.8 to 1.4), aggressiveness (9.6 to 1.0), and sprightfulness (16.0 to 21.6). All CG outcomes were statistically insignificant. In the IG outcomes, only depression, fatigue, and confusion scores in the POMS were statistically insignificant. These results show that RAS in conjunction with physical therapy can improve the gait and the psychological parameters of adults with CP (Efraimidou et al.).

Therapeutic Instrumental Music Performance

TIMP is the NMT technique that uses different musical instruments to help exercise the motor function of any deficient parts of the body to regain functional patterns of movement. The rhythmical nature of music while playing an instrument is conducive to triggering motor function and synchronizing movement and rhythm. Furthermore, the physicality of playing an instrument allows the patient to stimulate, train, and reproduce repetitive movement patterns used in daily life. This allows patients to more easily and more efficiently plan, program, and execute voluntary movements in their daily lives (Mertel).

2021 Study by Bilinc Karatekin and Afitap Icagasioglu

In 2021, Karatekin and Icagasioglu, investigated the effects of TIMP on motor function in the upper extremities of adolescents who were diagnosed with a form of spastic CP. The study took place in the Pediatric Disability Clinic at Istanbul Medeniyet University. A total of 9 participants, aged between 11 to 18 years old, all with CP with upper limb involvement were selected for this study. Pre-treatment and post-treatment measurements included hand functions measured using the Manual Ability Classification System (MACS), gross motor functions measured with the wooden box block test (BBT), upper extremity function and finger dexterity with the 9-hole peg test (NHPT), grip strength with the Jamar hand dynamometer, and their key

pressing force (only 5 of the participants) with the Cubase Midi program. Training sessions of 40 minutes were held twice a week for 3 months. In these sessions the participants began by playing notes of their choice on the piano which was followed by playing a pentatonic scale. After that they would play with two hands by pressing corresponding keys on the two separate hands simultaneously. Throughout the 3-month treatment the participants would also choose a musical piece and aim to play it by the end of the treatment.

Results

The results of these treatments showed many improvements across most of the tests. The MACS score after treatment was statistically significant with an improvement from around 2.89 to around 2.44. BBT scores were statistically significant with improvements in the right hand from around 23.11 box/min to around 28.33 box/min and in the left hand from around 28.56 box/min to around 32.78 box/min. NHPT scores were statistically significant with improvements in the right hand from around 171.37 seconds to around 125.88 seconds and from around 60.04 seconds to around 55.42 seconds in the left hand. Grip strengths in the Jamar hand dynamometer tests were statistically significant with improvements in the right hand from around 13.11 lbs to around 18.22 lbs and around 18.33 lbs to around 24.11 lbs in the left. In the MIDI test, only certain fingers had statistically significant increases which were in the right thumb (83.89 to 87.67), the right pointer finger (89.00 to 92.89), the right middle finger (80.78 to 90.89), the right ring finger (75.11 to 82.56), the left pointer finger (89.33 to 95.78), the left ring finger (83.33 to 89.56), and the left pinky finger (71.11 to 75.78). The results of this study indicate that TIMP is effective in improving the upper motor functions and the grip strengths of adolescent patients with CP (Karatekin and Icagasioglu).

2013 Study by Hyun Chong et al.

In 2013, Chong et al. investigated the effects of TIMP for fine motor exercises in adults with CP in Korea. The participants were patients at a rehabilitation ward in a university medical center. There were 5 female participants, aged 20 to 39, with either dyskinetic CP or a form of spastic CP. Pre-treatment and post-treatment, all 5 individuals had their hand functions measured using MIDI, a software program which measures key pressing force. Training sessions of 30 minutes each were held at least once a week for up to 9 weeks. The sessions involved repeated pressing (playing individual notes repeatedly) first, which was followed by sequential playing (playing ascending and descending adjacent keys) and then ended with simultaneous pressing (playing up to 3 keys at once with different fingers).

Results

While the mean scores of the individual fingers did increase there were no statistically significant differences in any of the pre and post scores. These results show that while TIMP may have the potential to improve fine motor skills in adults with CP, but unlike the later study by

Karatekin & Icagasioglu this study did not demonstrate statistically significant changes (Chong et al.).

Patterned Sensory Enhancement

PSE is the NMT technique that uses rhythmic, melodic, harmonic, and dynamic-acoustical elements of music to improve the motor function of movements that aren't rhythmical by nature. PSE uses musical patterns to combine separate movements together into simpler movement patterns and more complex movement sequences that can be used in daily life. Additionally, PSE works to improve physical strength, endurance, balance, posture, coordination, range of motion, and motor skills (Thaut).

2013 Study by Tze-Hsuan Wang et al.

In 2013, Wang et al. investigated the effects of PSE treatment along with resistance exercises on children with spastic cerebral palsy. The study used 36 children, all with spastic diplegia and aged 5 to 13 years old. The children were then split into two groups, based on their age. One group had participants aged 5 to 8.5 years and the other had participants aged 8.6 to 13 years. Measurements were taken post intervention 6 weeks in and 12 weeks in. The measurements made were gross motor capacity using the Gross Motor function Measure (GMFM), daily mobility and self-care using Pediatric Evaluation of Disability Inventory (PEDI), strength with the 1-repetition maximum (1-RM) of the STS tests and walking speed using the 10-meter walking test (10 MWT). Both groups went through the treatment where they participated in sit-to-stand (STS) resistance exercises. These exercises involved the participant putting on a weighted vest and standing up from a sitting position on a chair. One group was also prescribed the PSE treatment where the participant would do the STS exercise while listening to their individualized PSE music. The tempo, meter, and other musical elements were adjusted as the intervention progressed.

Results

The results in the PSE group showed statistically significant improvements compared to the non-PSE group. GMFM scores were statistically significant, compared to the non-PSE group, in the PSE group with an increase from a 71.1 to a 75.7, a 2.3 greater increase than the non-PSE group. The PEDI scores in the functional skills category were significantly significant, compared to the group without PSE, in the PSE group with an increase from 73.1 to 76.3, a 2.2 higher increase than the non-PSE group. However, there was statistical insignificance between the PSE group and the non-PSE group in outcomes of the mobility functional skills, mobility caregiver assistance, and self-care functional skills categories of the PEDI scores. There also was statistical insignificance between the PSE group and the non-PSE group in the outcomes of the 1-RM tests. From the results of this study, we can conclude that PSE has promise in improving the gross motor capacity and some functional skills in children with spastic CP (Wang et al.).

2011 Study by Yi-Chun Peng et al.

In 2011, Peng et al. investigated the immediate and the continual effects of PSE treatment on children with spastic CP in conjunction with loaded sit-to-stand (LSTS) treatment. The study used 23 participants, all aged 5 to 12 years old, who were randomly split into a PSE treated group and a non-PSE control group. Pre-intervention and post-intervention the researchers measured the 1-repetition maximum load for the LSTS movement (1-RM) for each participant in these groups. The motion of the center of mass (COM) of the participants was also measured during the treatments using infrared retroreflective markers and a 7-camera motion analysis system. LSTS involves the participant standing up from a chair with a weighted vest, then touching a rope at body height, sitting back down, and then immediately standing back up. The LSTS movement was split into two phases: phase 1 and phase 2. Phase 1 was the first half of the movement which starts with the initial movement of the body and ends when the participant is off the seat. Phase 2 was the last half which starts when the first phase ends and ends when the participant is at full standing posture. The PSE treatment music was individualized to the participants by a music therapist who made sure that the music had a 32-beat prelude, a tempo based on 95% of the time of phase 1, and the meter was either 2/4, 3/4, 4/4, or 6/4 beat. During the treatment the PSE music was used as cues for the change from phase 1 to 2 using the ascent or descent of the melodic lines, the loudness of the music, and the harmony of the music. Treatment involved LSTS movements with either no PSE music or with PSE music. In the PSE group the individuals were asked to do additional repetitions without the PSE music after they completed the PSE guided repetitions to see the continual effects of PSE.

Results

The results showed statistical significance in many of the tests. In the immediate measurements knee peak extensor power per limb was statistically significant in the PSE group compared to the control group with the PSE group being 0.12 Watt/N/m and the control group being 0.1 Watt/N/m. In the immediate measurements of motor control, the PSE group was statistically significant from the control group in the jerk index in the PSE group being around 4.74 while it was 6.03 in the control group and with the movement time in the PSE group being 1.92 seconds while it was 2.33 seconds in the control group. In the continued measurements, the continual group showed statistical significance from the control group in multiple parameters which were peak knee extensor power per limb at 0.11 Watt/N/m compared to 0.1 Watt/N/m, total extensor power per limb at 0.27 watt/N/m compared to 0.27 Watt/N/m, normalized jerk index at 3.06 compared to 6.03, and finally the movement time at 2.04 seconds compared to 2.33 seconds. From this study we can see that PSE can be effective in the short-term and in the following repetitions of treatment at increasing the knee peak extensor power and decreasing jerkiness in children with spastic CP (Peng et al.).

Discussion

The evidence reviewed in this paper suggests that NMT can improve both the physical and psychological wellbeing of individuals with CP. Across both studies, RAS consistently demonstrated effectiveness in improving gait related parameters such as stride length, velocity, and symmetry, especially when in conjunction with physical therapy (Kwak; Efraimidou et al.). Improvements were also noted in certain psychological measures such as mood and self-esteem (Efraimidou et al.). Overall, RAS has proved to be an effective treatment for managing the physical and psychological effects of CP.

TIMP interventions showed promising results in improving motor function in the upper extremities, particularly among adolescents with upper extremity affecting CP (Karatekin and Icagasioglu). However, findings were mixed overall due to the lack of statistically significant outcomes in the earlier study on adults (Chong et al.). These discrepancies may reflect the age differences of participants between the two studies, or limitations in treatment duration. Overall, TIMP shows promise as a possible treatment for managing the physical effects of CP.

PSE interventions, when integrated with resistance-based activities such as LSTS training, consistently demonstrated significant improvements in gross motor function, knee extensor power, and reductions in movement jerkiness in children with spastic CP (Wang et al.; Peng et al.). Overall, PSE has proved to be an effective treatment for managing the physical effects of CP.

One of the most valuable attributes of these NMT techniques is their ease of integration into existing physical treatments. This is most highlighted in RAS and PSE as they can easily be implemented into physical therapy treatments that patients are already using or commonly use. Therefore, implementing these NMT techniques into a patient's treatment plan could allow patients to gain additional benefits from their treatments in a relatively easy way.

However, it should be noted that there are limitations to these studies that must be acknowledged. All the studies reviewed had small sample sizes ranging from 5 to 36 participants. Intervention durations were short with the longest only lasting 3 months. Additionally, variations in research design, interventions, and outcome measures make comparisons between studies difficult. There is also a lack of studies examining the standalone effects of NMT, as nearly all the studies implemented NMT interventions in conjunction with traditional physical therapy.

Given these limitations, further research is necessary. Future research should include larger and more diverse sample populations, longer treatment durations, and should investigate further into the isolated effects of NMT techniques.

Conclusion

This review explored the potential of the NMT techniques RAS, TIMP, and PSE as treatments for CP patients. The findings suggest that these techniques can be an effective treatment to certain aspects of CP. RAS is effective in helping the patient achieve a more normalized gait along with improving their psychological well-being, TIMP is effective at

improving the motor skills of the upper extremities, and PSE is effective at improving the motor control of the lower extremities.

Although CP remains incurable, therapies should seek to mitigate the physical and psychological effects of CP to enhance the quality of life and the safety of patients. NMT, particularly when integrated into existing treatment programs, offers a promising, low-barrier complement to traditional treatments.

While current evidence is encouraging, broader and more rigorous research is needed to fully assess the efficacy and scope of these interventions. With further research, NMT could become a standard complement to the care of individuals with CP.

Funding and Conflict of Interest

No funding was received for this study. The author declares no conflict of interest.

Works Cited

- American Music Therapy Association. "What Is Music Therapy?" American Music Therapy Association, 2005, <https://www.musictherapy.org/about/musictherapy/>.
- Chong, H. J., et al. "Finger Exercise with Keyboard Playing in Adults with Cerebral Palsy: A Preliminary Study." *Journal of Exercise Rehabilitation*, vol. 9, no. 4, 2013, pp. 420–425. <https://doi.org/10.12965/jer.130050>.
- Dogruoz Karatekin, B., and A. Icagasioglu. "The Effect of Therapeutic Instrumental Music Performance Method on Upper Extremity Functions in Adolescent Cerebral Palsy." *Acta Neurologica Belgica*, vol. 121, no. 5, pp. 1179–1189, <https://doi.org/10.1007/s13760-021-01618-0>.
- Efraimidou, V., et al. "The Effect of a Music and Movement Program on Gait, Balance, and Psychological Parameters of Adults with Cerebral Palsy." *Journal of Physical Education and Sport*, vol. 16, no. 4, 2016, pp. 1357–1364, <https://doi.org/10.7752/jpes.2016.04217>.
- Graham, H. K., et al. "Musculoskeletal Pathology in Cerebral Palsy: A Classification System and Reliability Study." *Children*, vol. 8, no. 3, 2021, p. 252, <https://doi.org/10.3390/children8030252>.
- Hallman-Cooper, J. L., and F. Rocha Cabrero. "Cerebral palsy." *StatPearls*, StatPearls Publishing, 2024, <https://www.ncbi.nlm.nih.gov/books/NBK538147/>.
- Jan, M. M. "Cerebral palsy: Comprehensive review and update." *Annals of Saudi Medicine*, vol. 26, no. 2, 2006, pp. 123–132, <https://doi.org/10.5144/0256-4947.2006.123>.
- Kern, P., and D. B. Tague. "Music Therapy Practice Status and Trends Worldwide: An International Survey Study." *Journal of Music Therapy*, vol. 54, no. 3, 2017, pp. 255–286, <https://doi.org/10.1093/jmt/thx011>.
- Krigger, K. W. "Cerebral Palsy: An Overview." *American Family Physician*, vol. 73, no. 1, 2006, pp. 91–100.
- Kwak, E. E. "Effect of rhythmic auditory stimulation on gait performance in children with spastic cerebral palsy." *Journal of Music Therapy*, vol. 44, no. 3, 2007, pp. 198–216, <https://doi.org/10.1093/jmt/44.3.198>.
- Mertel, K. "Therapeutic instrumental music performance (TIMP)®." *Handbook of Neurologic Music Therapy*, edited by M.H. Thaut and V. Hömberg, 2nd ed., Oxford University Press, 2025, pp. 128–152, <https://doi.org/10.1093/oso/9780192844491.003.0010>.
- Papavasiliou, A. S. "Management of Motor Problems in Cerebral Palsy: A Critical Update for the Clinician." *European Journal of Paediatric Neurology*, vol. 13, no. 5, 2009, pp. 387–396, <https://doi.org/10.1016/j.ejpn.2008.07.009>.
- Patel, D. R., et al. "Cerebral Palsy in Children: A Clinical Overview." *Translational Pediatrics*, vol. 9, suppl. 1, 2020, pp. S125–S135, <https://doi.org/10.21037/tp.2020.01.01>.
- Peng, Y.-C., et al. "Immediate Effects of Therapeutic Music on Loaded Sit-to-Stand Movement in Children with Spastic Diplegia." *Gait & Posture*, vol. 33, no. 2, 2011, pp. 274–278, <https://doi.org/10.1016/j.gaitpost.2010.11.020>.

- Shamsoddini, A., et al. "Management of Spasticity in Children with Cerebral Palsy." *Iranian Journal of Pediatrics*, vol. 24, no. 4, 2014, pp. 345–351.
- Thaut, C. P. "Patterned Sensory Enhancement (PSE)®." *Handbook of Neurologic Music Therapy*, edited by M. H. Thaut and V. Hömberg, 2nd ed., Oxford University Press, 2025, pp. 115–127, <https://doi.org/10.1093/oso/9780192844491.003.0009>.
- Thaut, C. P., and R. Rice. "Rhythmic Auditory Stimulation (RAS)." *Handbook of Neurologic Music Therapy*, edited by M. H. Thaut and V. Hömberg, 2nd ed., Oxford University Press, 2025, pp. 99–114, <https://doi.org/10.1093/oso/9780192844491.003.0008>.
- Thaut, M. H., and V. Hömberg, editors. *Handbook of Neurologic Music Therapy*. 2nd ed., Oxford University Press, 2025, <https://books.google.com/books?hl=en&lr=&id=2QdIEQAAQBAJ>.
- Tugui, R. D., and D. Antonescu. "Cerebral Palsy Gait, Clinical Importance." *Maedica*, vol. 8, no. 4, 2013, pp. 388–393, <https://pubmed.ncbi.nlm.nih.gov/24790675/>.
- Vinolo-Gil, M. J., et al. "Effects of the Combination of Music Therapy and Physiotherapy in the Improvement of Motor Function in Cerebral Palsy: A Challenge for Research." *Children*, vol. 8, no. 10, 2021, p. 868, <https://doi.org/10.3390/children8100868>.
- Vitrikas, K., H. Dalton, and D. Breish. "Cerebral Palsy: An Overview." *American Family Physician*, vol. 101, no. 4, 2020, pp. 213–220.
- Wang, T.-H., et al. "A Home-Based Program Using Patterned Sensory Enhancement Improves Resistance Exercise Effects for Children with Cerebral Palsy: A Randomized Controlled Trial." *Neurorehabilitation and Neural Repair*, vol. 27, no. 8, 2013, pp. 684–694, <https://doi.org/10.1177/1545968313491001>.

To What Extent Did Ideology and Propaganda Contribute to Mao's Rise to Power?

By Xinyou Chen

Introduction

The Qing dynasty collapsed on October 10th of 1911 due to the Wuchang Uprising after 268 years of ruling since 1644¹⁶. The collapse of the Qing dynasty was not merely the end of a dynasty, but the end of China's monarchy, which had persisted for over 3000 years. This left China with a power vacuum, leading to numerous attempts from different forces to fill this vacuum, culminating in the Chinese Communist Party (CCP)'s ultimate seizure of power led by Mao Zedong. Although Mao's use of ideology and propaganda was a contributing factor to his seizure of power, it was the Guomindang (GMD)'s failure to maintain popular support and effective governance to fill the power vacuum and also the impact of foreign invasions that are the most significant factors that led to Mao's usurpation of power. This is because the GMD's failure to fill the power vacuum created an opportunity for Mao's spread of ideology and propaganda which exploited the masses' suffering and frustration caused by foreign invasions

While this essay examines the different levels of significance that the three contributing factors have on Mao's rise to power, it is important to recognize that it is ultimately the interplay of the three factors that led to the CCP's seizure of control. All three factors are interdependent, and the absence of one can diminish the effectiveness of another. Different historians demonstrate varying perspectives on the importance of the three factors. This essay's claim comes from the evaluation of different sources, which comprehensively consist of opposing viewpoints and various historians. Immanuel Hsu, John K. Fairbank, and Chalmers A. Johnson are historians supporting the structuralist argument, placing greater significance on the conditions Mao stepped into. In *Peasant Nationalism and Communist Power*, Johnson emphasized the key role of the Sino-Japanese War in escalating the pre-existing discontent among the peasantry in order to fuel a large-scale peasant revolt.¹⁷ By the same token, Bianco also emphasized the condition of Chinese society. However, Bianco's argument is nuanced as it discusses China's social condition in the context of how it bolstered Mao's propaganda and ideology. Thus, instead of simply supporting that one factor is more important than the other, Bianco supports the view that the social conditions only came to effect with Mao's ideology and propaganda. On the contrary, historians including Mark Seldon support the intentionalist viewpoint. In *China in Revolution: The Yen'an Way Revisited*, Selden credited Mao's progressive reforms and effective ideology for the CCP's rise.¹⁸ While these major historians' work and other secondary sources allowed for a broad understanding of Mao's rise to power, it is important to acknowledge the limited access to primary sources from the CCP and GMD. The primary

¹⁶ *Timeline of China's modern history*. Chicago Public Library. (n.d.).

¹⁷ Johnson, C. A. (1963). *Peasant nationalism and Communist power: The emergence of revolutionary China, 1937-1945*. Stanford U.P. p.4

¹⁸ China in revolution: The yen'an way revisited (review). (1998). *China Review International*, 5(1), 240–242. <https://doi.org/10.1353/cri.1998.0042>

sources used in this essay include Mao's "Draft Resolution of the Second Congress of County Party Representatives of the Xiang-Gan [Hunan-Jiangxi] Border Area" written in 1928. While these sources are valuable for historians to understand Mao's initial thought on ideology, there are limited primary sources that present the harmful side of Mao's propaganda as he gradually gains more control over state press and media.

This essay claims that while Mao's use of ideology and propaganda and Guomindang mistakes and systemic weaknesses contributed to his rise to power, foreign invasions and the power vacuum they created were the most important factors contributing to Mao's rise to power. Foreign invasions not only cultivated anti-imperialism and nationalism but also exposed the GMD's deeper systemic problems and intensified peasant suffering under the GMD. Ideology and propaganda is the least important factor in Mao's rise to power, as the GMD's mistakes and foreign invasion together created a condition under which the former became possible.

Historical Context

After the fall of the Qing dynasty and the abdication of Manchu emperor Puyi, Yuan Shikai came to power without any revolutionary or influential policies, while Sun Yatsen became the provisional president of the Republic of China¹⁹. During the time when China was in a vulnerable transitional stage from the end of the monarchy to the start of the republic, many foreign forces invaded and defeated China, such as the Eight-Nation Alliance, the second opium war, and the First Sino-Japanese War. After Yuan's death in 1916, the national power collapsed, prompting the start of the Warlord Era in China. Warlordism is characterized by fragmented regional leadership. Regional rulers consisted of commanders of the Yuan's Beiyang army with considerable military power. Due to Warlordism's inherent division and instability, the Warlord Era was one of economic stagnation and social repression, causing the increased suffering of the peasants. Although Chiang Kaishek, the new leader of the GMD, was able to end Warlordism through the northern expedition with the First United Front in alliance with the CCP, Chiang neglected the rural peasants' plight and struggles, and ended the First United Front subsequently. The end of the First United Front led to the White Terror campaign by the GMD which marked the first stage of the Civil War between the CCP and GMD, including a series of events such as later interrupted by the Japanese invasion and the formation of the Second United Front. The collapse of the Second United Front followed the second stage of the Civil War, which resulted in Mao's victory by 1949.

Section 1: Foreign invasion and the resulting power vacuum

Foreign invasion created a condition that is important to Mao's rise to power through cultivating anti-imperialist sentiment, creating economic instability, and most importantly, producing a power vacuum that the Chinese Communist Party (CCP) effectively exploited. The early 20th century saw numerous foreign invasions before the Second Sino-Japanese War (1937)

¹⁹ *Timeline of China's modern history*. Chicago Public Library. (n.d.). <https://www.chipublib.org/timeline-of-chinas-modern-history/>

that targeted China's vulnerability and sparked widespread anti-imperialist sentiment and nationalism, especially among the intellectuals. This is evident through the May Fourth Movement in 1919 that emerged as a response to the Treaty of Versailles, which was seen as a betrayal of Chinese interests.²⁰ The May Fourth movement was an anti-imperialist protest from students of Peking University, which Mao also took part in as a librarian in Peking University at the time. The awakening of nationalism is significant as it generated an emotional need among the Chinese masses, particularly students and urban youth, for an ideology that emphasizes dignity, sovereignty, and independence, which the CCP can take advantage of. It also led to the creation of the Communist Party, where students were convinced that the only way forward for China was via communism.

Furthermore, foreign invasion and intervention led to the fragmentation of the ruling class. From the ineffective presidency of Yuan Shikai to Sun Yatsen's failed attempts to fill the power vacuum, the Chinese masses have never been unified by a singular, effective leader. This situation is worsened by the rise of Warlords. Yuan Shikai's control of the military gave him the pathway to the presidency.²¹ However, the military strength of these military commanders remained after Yuan's death in 1916. The Warlords were regional rulers who were previously a part of the Yuan's Beiyang army. Peasants in rural China were victims of the Warlord Era, and the warlords "sought to increase their power by increasing the size of their armies", leading to conscription and coercion. Additionally, Warlords sought wealth from the ordinary people, particularly the peasants in rural areas, through exploitation, corruption, and banditry, which had dire effects on the peasants. Mao wrote in his own work *Draft Resolution of the Second Congress of County Party Representatives of the Xiang-Gan [Hunan-Jiangxi] Border Area*: "China's localized and divided political economy makes it impossible for the ruling class to sustain unified action. Mao also emphasized that "the Chinese audience had had ample experience of the chronic contention among various powerholders".²² These viewpoints are significant because the warlord era not only created distress in the peasant population but also demonstrated that there was a future opportunity to mobilize the peasants by fulfilling their needs and addressing their continual struggles. While Mao did not fully understand the potential of peasant distress at the time, he was able to recognize the importance of peasant mobilization after the First United Front. This is significant because the fragmentation of China due to the warlord rule made it harder for a single leader to step in and take complete control of China, including Chiang.

Another example of the impact of war on Mao's rise to power was due to the Second Sino-Japanese War (1937-1945), as it led to an unstable economic condition that created political instability, which Mao was able to exploit. The war between China and Japan led to the GMD purchasing weapons in huge quantities leading to an insufficiency in the GMD's budget. This led

²⁰ Ibid.

²¹ Alphahis. (2025, March 2). *The warlord era*. Chinese Revolution. https://alphahistory.com/chineserevolution/warlord-era/#Rise_of_the_warlords

²² Womack, B. (2018, September 19). *3Rural revolution: 1927–1931*. Project MUSE. https://muse.jhu.edu/pub/5/oa_monograph/chapter/2179697

to inflation, causing the rising prices of necessities and economic instability.²³ This is significant because it had a great impact on the middle class and peasants, fueling discontent towards the GMD government. This discontent caused political instability as the GMD became the scapegoat of the common citizens' struggle. This is significant because the Second Sino-Japanese War created an unstable political condition and a withering of trust peasants have towards the GMD. Historian Immanuel Hsu attributes Mao's rise and the GMD's fall mainly to the impact of the Sino-Japanese War. He described it as "the single most important near cause for the downfall of the Nationalists"²⁴. Hsu supports the structuralist perspective that conditions and the CCP's ability to fill the power vacuum as the greatest conditions for Mao's rise to power. However, other historians such as Lucien Bianco support the intentionalist argument and Mao's methods that led to the CCP's success. He focuses on their adaptation to guerrilla warfare in the second stage of the Civil War and how they took advantage of the peasants through Mao's ideology to aid as the greater reason.²⁵ However, this disregards Hsu's argument and overlooks the importance of the Second Sino-Japanese War. This is significant because the power vacuum caused by foreign invasion gave the CCP an opportunity to gain traction, meanwhile exposing the weakness in the GMD leadership and their continued corruption. Mao took advantage of this during the Civil War where he was ultimately victorious. Foreign invasion was a prerequisite for Mao to effectively exercise his ideology, reforms, and propaganda, and thus, the most important factor contributing to Mao's rise to power.

Section 2: Mistakes made by the Guomindang(GMD)

While not as significant as foreign invasion and the subsequent power vacuum it generated, the Guomindang also made fatal mistakes that led to the loss of support from the middle class, intellectuals, and peasants -- demographic groups whose support was quintessential for the success of any political party in China. Firstly, the Guomindang failed to capitalize on the anti-imperialist sentiment that emerged among the intellectuals and peasants. This failure largely stemmed from Chiang Kai-shek's prioritization of the suppression of the Chinese Communist Party (CCP) over resisting the Japanese invasion. Chiang stated that "The Japanese are a disease of the skin, the communists are a disease of the heart."²⁶ This statement illustrates the GMD's strategic focus on eliminating internal opposition rather than defending the nation against foreign aggression. This is specifically evident through the Nanking Massacre, when the GMD troops abandoned the city of Nanking and left millions of civilians to the atrocities committed by the Japanese army. The GMD's lack of investment in fighting the Japanese is significant as it fueled

²³ Sun, H. (2021). *(pdf) interplay between nationalist weaknesses and communist strengths in the outcome of the second stage of the Chinese Civil War (1946–1949)*. Scholarly Review Journal. https://www.researchgate.net/publication/382837354_Interplay_Between_Nationalist_Weaknesses_and_Communist_Strengths_in_the_Outcome_of_the_Second_Stage_of_the_Chinese_Civil_War_1946-1949

²⁴ Hsü, I. C. Y. (1970). *The rise of modern China*. Oxford University Press. p.213

²⁵ Bianco, L. (1993a). *Origins of the Chinese Revolution, 1915-1949*. Stanford University Press

²⁶ Theodore H. White *In Search of History* (1978)

growing resentment among the peasants who suffered the most under Japanese occupation. This also drove discontent in intellectuals, whose nationalist aspirations had been sparked after witnessing China's constant defeat by foreign forces in the late 19th century and early 20th century, along with the May Fourth Movement. Additionally, this gave the Communists an opportunity to demonstrate their determination in defeating the Japanese troops and to promote themselves as the true defenders of China's sovereignty. This narrative appealed to both peasants' suffering and the intellectuals' ideological demands that centered on anti-imperialism²⁷. Mao further capitalized on the GMD's mistake by portraying the GMD as capitulators to imperialism. The GMD was supported by the USA during the Civil War due to the US' anti-communist position. Mao's commanders publicly display the "made in USA" phrases on captured GMD weapons in CCP-controlled areas and would inscribe this phrase if it was missing.²⁸

Additionally, the GMD also failed to implement effective socio-economic policies that appeal to the middle class. Moreover, the GMD completely neglected the needs of the peasant population in China, leading to the suffering of the peasant masses. While some historians, such as Parks M. Coble, attribute the Nationalist government's economic challenges to external factors and emphasize the limited capacity of the GMD to address peasant hardships.²⁹ Bianco supports the argument that it is "relatively easy for GMD to redistribute land through reforms."³⁰ The latter is supported by the statistics that "from 1935-36, only 3.7% of the national budget was allocated to rural reconstruction,"³¹ reflecting the GMD's failure to prioritize the population that constituted the majority of Chinese. This indicates that GMD's internal policy failures played a more significant role in alienating the peasantry, giving rise to Mao's propaganda that specifically targeted the alienated peasantry which largely contributed to his rise to power.

Lastly, GMD's corruption and ineffective leadership during the second stage of the Civil War exacerbated its declining popular support and legitimacy. GMD generals were selected on the basis of their loyalty to Chiang instead of their military intelligence and competency. "Chiang repeatedly recalled and replaced commanders, resulting in a discontinuity in leadership that crippled any possible Nationalist initiative"³² The lack of fairness and transparency in the GMD's system led to ineffective attacks during the second stage of the Civil War. This gave rise to Mao, who selected commanders based on their military vision and ensured continuity in leadership. Corruption in the GMD generals was prevalent, including stealing and selling government properties. While the commanders were granted huge power, the GMD soldiers

²⁷ Shuja, S. M. (1976). WESTERN INVOLVEMENT IN CHINA (1557— 1949): CHALLENGE TO CHINESE SOVEREIGNTY. p.6.

²⁸ Bianco, L. (1993a). *Origins of the Chinese Revolution, 1915-1949*. Stanford University Press. p.186

²⁹ The Collapse of Nationalist China: How Chiang Kai-shek Lost China's Civil War By Parks M. Coble. Cambridge: Cambridge University Press

³⁰ Bianco, L. (1993a). *Origins of the Chinese Revolution, 1915-1949*. Stanford University Press. p.110

³¹ Ibid.

³² Fairbank, J. K. M. 2006, "Japan's invasion in the 1930s played a major role in reinvigorating 'China's new nationalism'" p. 124

received inadequate pay and lacked a clear purpose in their fighting, ultimately making them demoralized.³³ An opposing viewpoint is supported by historian Van de Ven, stating “the intelligent Communist strategy of fighting only when success cultivated a feeling of victory and stimulated a spirit of boldness that contrasted vividly with the defensive mentality of the GMD”³⁴ Van de Ven emphasizes the importance of Mao’s strategies and the strong sense of purpose he instilled in the soldiers, turning the GMD’s weakness into a facilitating factor for the CCP’s success. However, Hsu places a greater significance to the dire conditions of the GMD: “Desertion and defection of entire units were common as soldiers were inadequately paid and fed; training was poor to nonexistent; discipline was bad; men did not know what they were fighting for, and thus saw no reason to fight.”³⁵ Hsu’s perspective highlights the lack of organization and cohesion in the GMD force, which provided the CCP with opportunities to win battles, which were leveraged to expand political influence and reinforce the morale and commitment of their troops. GMD’s internal system is highly dictatorial, which not only disappoints soldiers and generals internally, but also led to the collapse of its outward claim to uphold democratic values, delegitimizing the party’s central ideology.³⁶ This shows GMD’s internal corruption is more important than Maoist ideology and propaganda because Mao’s ideology and propaganda towards the intellectuals and peasants targeted their existing disillusionment with the Nationalists, offering them renewed hope. However, this is not as significant as the power vacuum created by foreign invasion. While the GMD’s systemic flaws did exist prior to the Sino Japanese War, they had not been exposed to the public to such an extent before the war. Foreign invasion catalyzed the long-standing discontent of the peasants, intensifying the frustration to finally erupt. Moreover, defeating a foreign power demanded a disciplined, unified, and organized military, which exposed GMD’s structural weaknesses in GMD’s leadership. These would have otherwise remained overlooked. Therefore, while not as significant as foreign invasion and the impact of war, the GMD’s mistakes that led to weaknesses in the political system that Mao exploited are more important than Mao’s ideology and propaganda.

Section 3: Ideology and propaganda

While both foreign invasion and mistakes made by the Guomindang contributed more to Mao’s rise to power, Mao’s skillful use of ideology and propaganda with a meaningful target population also resulted in the Chinese Communist Party’s (CCP)’ rise to power.

³³ Sun, H. (2021). *(pdf) interplay between nationalist weaknesses and communist strengths in the outcome of the second stage of the Chinese Civil War (1946–1949)*. Scholarly Review Journal. https://www.researchgate.net/publication/382837354_Interplay_Between_Nationalist_Weaknesses_and_Communist_Strengths_in_the_Outcome_of_the_Second_Stage_of_the_Chinese_Civil_War_1946-1949

³⁴ (Van De Ven, 2017, p. 86)

³⁵ Hsü, I. C. Y. (1970). *The rise of modern China*. Oxford University Press. p.288

³⁶ Sheridan, J. E. (1975). *China in disintegration: The Republican era in Chinese history, 1912-1949* Free Press. p.23

Mao's success in using ideology and propaganda is attributed to his flexibility in adapting the Marxist ideology to China's social context in the 20th century. While the original Marxist idea centers around the readiness of industrial working classes to achieve revolution, Mao's communism was not limited by such an idea. Mao's communism in both stages of China's Civil War was based on adjusting and shifting this idea to the rural areas and peasants' context of China.³⁷

In Mao's 1928 writing "Draft Resolution of the Second Congress of County Party Representatives of the Xiang-Gan [Hunan-Jiangxi] Border Area," he outlined the five conditions that could lead to the rise of the CCP, upon which he based the Maoist ideology, including first, the fragmentation of the ruling class due to warlordism and imperialism. Second, the mass participation in prior revolutionary activity to build revolutionary consciousness. Third, a link to the broader national revolutionary movement, a strong red army, and lastly the presence of effective leadership and policies inside the party³⁸. While outlined as conditions, each condition required active strategic effort made by the CCP to turn these conditions to their advantage and work to their advantage. Both capitalized on the incontinuity of the ruling class and building the masses' revolutionary consciousness directed Mao to put effort into mobilizing China's peasantry. The peasantry's large population in China existed alongside the Chinese Communist Party's limited urban influence. Thus, the peasants' long-neglected but deeply rooted dissatisfaction was an emotional and ideological potential that the CCP could strategically exploit and mobilize. This is not as important as the conditions created by foreign invasions and the GMD's mistakes, however, since Mao's ideology and the quintessential factors to succeed as outlined in "Resolution" are largely contingent on taking advantage of the existing conditions of China, which he based the subsequent spread of propaganda and reforms on. A limitation to the third condition in this writing is that at the time, Mao had yet to realize that success in villages did not have to engage with the broader proletarian demographic's revolution to be successful. Mao successfully seized power through a prolonged peasant revolution.

Mao executed his ideology by gaining the support of Chinese peasants through reforms such as those of land ownership and taxes, and propaganda and the People's Liberation Army's respect and care, representing the CCP's common touch and the cultivation of peasant nationalism. As demonstrated in Appendix Fig. 1, the conduct of the People's Liberation Army solidified the positive image of Mao and the CCP. The PLA conducted careful military advances without disturbing the lives of peasants in contrast to the GMD's military forces, which were often seen looting and abusing local populations. "To millions of peasants, soldiers, and civilians alike, the Red Army (later the PLA) brought the immediate promise of a new existence, of liberation from all the evils of the old society"³⁹. The PLA's action-oriented propaganda's effectiveness is shown when recruitment of the PLA increased to 1.6 million from June 1946 to

³⁷ Womack, B. (2018, September 19). *3Rural revolution: 1927–1931*. Project MUSE. https://muse.jhu.edu/pub/5/oa_monograph/chapter/2179697

³⁸ Ibid.

³⁹ Bianco, L. (1993a). *Origins of the Chinese Revolution, 1915-1949*. Stanford University Press p.190

June 1948⁴⁰. Ultimately, the PLA's action garnered major support and was instrumental in the guerrilla warfare requiring cooperation from the peasantry. This is significant as the PLA's discipline and respectful conduct not only successfully differentiated themselves from the GMD and gave hope to China's depressed peasant society, but also became a form of living propaganda, embodying the values the CCP sought to promote. This shows ideology and propaganda are less important compared to opportunities GMD missed and the foreign invasion, as the latter two factors gave Mao the opportunity to capitalize on the peasants' discontent. Additionally, Mao's need to cultivate peasants' nationalism justified the chosen base area of Jiangxi as it is an isolated and socio-economically feudal society. Mao led peasant revolutions, such as the Autumn Harvest Uprising⁴¹, to promote resentment towards the GMD as a scapegoat of their socio-economic plight. This is significant as it cultivated an awareness for the peasantry to fight for their own class, meanwhile positioning the CCP as the force most capable and willing to help them achieve victory in the class struggle. Furthermore, Mao also implemented successful socio-economic reforms that were closely related to his ideology. Before the policies of Mao, peasants in rural China suffered silently from unequal land distribution, taxes, rent, and oppression by landlords (Sun, 2021). Socially, Mao promulgated the *Outline Land Law*, abolishing the land ownership of all landlords to distribute land equally among peasants. Economically, Mao made reforms on land interest rates, land tax, and land rent paid by the peasants to ease their financial burden.⁴² It is significant that Mao's economic policies are closely tied to the social policies, and his ideology is rooted in revolts from the peasants, as the success of these economic policies ends up strengthening the peasants' trust in the CCP's general ideology, not just limited to outcomes of any specific policy. Historian Mark Selden argues that Mao did not just step into a power vacuum but was able to win the enthusiastic support of the majority of the people in CCP-controlled areas by creating viable alternatives to decaying institutions.⁴³ However, Chalmers Johnson holds the belief that it was after the outbreak of the war that "The politically illiterate masses of China were awakened... wartime conditions made them receptive to a new kind of appeal... the defense of the fatherland". (Peasant nationalism and communist power."⁴⁴ Johnson positions foreign invasions as a force that challenged the peasants' security and intensified their struggle, giving rise to mass peasant revolution. This shows that Mao's ideology and propaganda is the last important factor that contributed to his rise to power as the effectiveness of his propaganda required the condition established by the GMD's failure to fill the power vacuum and the impact of war.

⁴⁰ Fairbank, J. K. M. 2006, "Japan's invasion in the 1930s played a major role in reinvigorating 'China's new nationalism'"

⁴¹ Womack, B. (2018, September 19). *3Rural revolution: 1927–1931*. Project MUSE. https://muse.jhu.edu/pub/5/oa_monograph/chapter/2179697

⁴² Ibid.

⁴³ Selden, M. (1998). China in revolution: The yenan way revisited (review). *China Review International*, 5(1), 240–242. <https://doi.org/10.1353/cri.1998.0042>

⁴⁴ Johnson, C. A. (1963). *Peasant nationalism and Communist power: The emergence of revolutionary China, 1937-1945*. Stanford U.P. p.2

Conclusion

Foreign invasion played the most significant role in causing Mao's rise to power in 1949. This is because these invasions not only fueled widespread anti-imperialist sentiment and strengthened nationalist movements but also revealed the Guomindang's underlying systemic weaknesses and exacerbated the suffering of the peasantry, which is the groundwork for Maoist ideology to gain traction. In contrast, ideology and propaganda were the least important contributors; their effectiveness was largely dependent on the conditions shaped by both foreign intervention and the Guomindang's failures. Without these enabling circumstances, Mao's ideological messaging would not have gained such widespread traction.

The claim of this essay is drawn from the evaluation of different historian's differing perspectives and the understanding of the historical debate on this topic. Despite intentionalist historians like Mark Selden crediting Mao's reforms and ideology as the primary drivers of the CCP's rise, Chalmers Johnson and Lucien Bianco highlight the importance of external conditions and their role in the ultimate success of Mao's ideology and propaganda.

As such, this essay contends that the rise of the CCP is a product of structural and geopolitical forces more than ideological innovation. Mao's propaganda may have built popular support, but it was the collapse of China's imperial structure, the failures of the Guomindang, and the chaos of foreign invasion that gave that message power.

Works Cited

- Alphahis. (2025, March 2). The warlord era. Chinese Revolution.
https://alphahistory.com/chineserevolution/warlord-era/#Rise_of_the_warlords
- Bianco, L. (1993a). *Origins of the Chinese Revolution, 1915-1949*. Stanford University Press.
- Bianco, L. (1993b). *Origins of the Chinese Revolution, 1915-1949*. Stanford University Press.
- Fairbank, J. K. M. 2006, “Japan’s invasion in the 1930s played a major role in reinvigorating ‘China’s new nationalism’”
- Hsü, I. C. Y. (1970). *The rise of modern China*. Oxford University Press. p.288
- Johnson, C. A. (1963). *Peasant nationalism and Communist power: The emergence of revolutionary China, 1937-1945*. Stanford U.P.
- Selden, M. (1998). China in revolution: The Yen-an way revisited (review). *China Review International*, 5(1), 240–242. <https://doi.org/10.1353/cri.1998.0042>
- Sheridan, J. E. (1975). *China in disintegration: The Republican era in Chinese history, 1912-1949*. Free Press. P.23
- Shuja, S. M. (1976). *WESTERN INVOLVEMENT IN CHINA (1557— 1949): CHALLENGE TO CHINESE SOVEREIGNTY*.
- Sun, H. (2021). (pdf) Interplay between nationalist weaknesses and communist strengths in the outcome of the second stage of the Chinese Civil War (1946–1949). *Scholarly Review Journal*.
https://www.researchgate.net/publication/382837354_Interplay_Between_Nationalist_Weaknesses_and_Communist_Strengths_in_the_Outcome_of_the_Second_Stage_of_the_Chinese_Civil_War_1946-1949
- Timeline of China’s modern history. Chicago Public Library. (n.d.).
<https://www.chipublib.org/timeline-of-chinas-modern-history/>
- Theodore H. White *In Search of History* (1978).
- Van de Ven H. (2017). *Triumph and tragedy in the emergence of the new china 1937-1*. Harvard University Press
- Womack, B. (2018, September 19). *Rural revolution: 1927–1931*. Project MUSE.
https://muse.jhu.edu/pub/5/oa_monograph/chapter/2179697

Appendix

Fig.1: the PLA sweeping the village they passed through



How Does the Brain's Perception of Episodic Memory Change over Time?

By Arjun Balasundaram

“The Neuroscience of Learning and Memory”
Robert E. Clark, Ph.D.

ABSTRACT

Episodic memory is important and vital because it determines how someone views themselves and the world around them(Pause et al., 2013). The objective of this review was to understand how episodic memory develops over time and to determine whether there is a difference in perceiving episodic memories that changes with age. The studies selected were found on the Science Direct database, and the papers included were focused on how episodic memory develops over time for different age groups. This paper found that episodic memory improves rapidly during childhood, with structural development in early childhood, strategic developments in middle childhood, and strategic refinements and improvements in overall efficiency in adolescence, peaking with young adults, and some aspects declining with old age. This concluded that childhood is the most important time for episodic memory development. Further research needs to be done to understand how people perceive their own episodic memories, and more work needs to be done on how the perception of a memory can, if possible, be objectively rated for a more concrete approach and understanding.

INTRODUCTION

Episodic memory is the ability to recall specific events from personal experiences, or, as Allen and Fortin (2013) described it, the ability to “‘mentally time travel’ to reexperience individual events”. It is one of the primary types of memory and falls under the broader category of declarative memory. Declarative memory refers to knowledge that can be consciously recalled and expressed in words (Pause et al. 2013). It is commonly divided into two subtypes: episodic memory, which stores personal experiences tied to specific times and places, and semantic memory, which stores general facts and knowledge about the world (Pause et al. 2013).

Declarative memory itself is a branch of long-term memory, which also includes non-declarative memory. Non-declarative memory is unconscious and not easily revealed through skills and habits (Pause et al. 2013).

Episodic memory, then, is unique in that it allows us to re-experience past events vividly (Pause et al. 2013), such as the first day of school or a friend’s birthday party. However, episodic memory’s accuracy can be influenced by factors such as how much attention was paid during the original event or how much time has passed since it occurred.

The importance of studying episodic memory lies in the central role of our daily lives. A better understanding of how we recall events can have wide-ranging applications, from improving eyewitness testimony in criminal justice to enhancing learning and retention in education. Research into episodic memory also has implications for mental health. For example, knowing that memories of frightening or anxiety-provoking experiences may sometimes be

exaggerated by the brain can help individuals reinterpret those experiences in less distressing ways (Pause et al. 2013)

Beyond practical applications, episodic memory is deeply tied to how people perceive themselves and the world. It shapes our identity and provides insight into the experiences of those around us, fostering empathy and social connection (Pause et al. 2013). Therefore, studying episodic memory not only advances our scientific understanding of how memory works but also deepens our understanding of human experience.

This review aimed to examine how episodic memory develops, summarize key findings in the field, and highlight areas where further research may be needed. By bringing these perspectives together, the goal was to provide a clearer picture of episodic memory as a fundamental part of the broader memory system.

METHODOLOGY

The literature search was conducted using the database Science Direct. The keywords used were "episodic memory development". To refine the results, the filter for article type was set to "research articles", while no restrictions were applied to publication type, subject area, access type, or publication year. The search and article retrieval were finalized on July 9th, 2025. To ensure relevance and quality, only peer-reviewed research articles published between 2000 and 2025 in English were included. Studies whose main focus was not the development of episodic memory across human life, or that worked with non-human subjects, were excluded from the paper collection.

The initial search identified 53 research papers. After following the inclusion and exclusion criteria, the number narrowed to nine papers, which seemed to focus on the understanding of episodic memory and its development in humans. These studies were organized according to the age groups investigated: early childhood, middle childhood, and late childhood/adolescence.

RESULTS

The final number of papers that followed all the inclusion criteria and were included in this review was nine, which together cover a significant portion of childhood as well as some aspects of adulthood (Table 1). These papers were divided into age groups, as most of the papers compared multiple age groups or focused on one age group in particular, so the highlighted findings in these papers will best relate to age groups.

Table 1. Titles, authors, year, and age group are represented in the selected papers.

Title	Author and year	Age group
White matter structural connectivity and episodic memory in early childhood	Ngo et al. 2017	4 - 6 Early childhood
Hippocampal functional connectivity and episodic	Riggin et al. 2016	4 - 6

memory in early childhood		Early childhood
Remembering the past to envision the future in middle childhood: Developmental linkages between prospection and episodic memory	Coughlin et al. 2014	5-9, and adults Early/middle childhood
Examining the development of memory for temporal context and its underlying neural processes using event-related potentials	Bettencourt et al. 2021	7-12, adults Middle childhood
Developmental differences in the neural correlates of relational encoding and recall in children: An event-related fMRI study	Güler and Thomas 2013	8-13 Middle childhood
Changes in familiarity and recollection across the lifespan: An ERP perspective	Friedman et al. 2010	9-10, 13-14, 20-30, 65-85 Adolescence
The underlying processes of episodic memory development: From a unique contribution of representation to the increasing use of semantic organization supported by cognitive control	Frick et al. 2022	6-13 Middle childhood
Children's and adults' memory for the order of events in a museum: A preliminary study about temporal memory and its development using photo-taking and event-related potentials	Pathman et al. 2014	7-10 Middle childhood
Episodic memory during middle childhood: What is developing?	Guo et al. 2024	6-12 Middle childhood

Both papers that cover early childhood analyze facets of episodic memory and how changes in childhood affect the development of episodic memory.

First, Ngo et al. (2017) studied the brain's structural connectivity, focusing on the white matter, a part of the brain that is believed to be important for cognition and basic neural functions (Filley and Fields, 2016). The study aimed to analyze the differences in macro and microstructure of white matter in the brain and the effects of age on white matter. For the analysis of structural connectivity of white matter, diffusion-weighted imaging was used. This method allows researchers to better understand the structure of white matter in the brain (Schilling et al. 2023). Analyzing the macrostructure includes the physical volume of the tract and the extent of white matter, which connects the hippocampus to other brain structures. The analysis of the microstructure studies the fractional anisotropy (FA), which serves as a reference to understand the integrity of white matter in the brain. A difference in the FA could serve to show a difference in the white matter of the brain. Ngo et al. 2017 aim to explain how such changes in the brain could affect episodic memory.

To study how differences in white matter caused by age affect episodic memory, the researchers used different methods to understand the children's episodic memory. The methods the researchers used were the children's memory scale (CMS) episodic memory tasks. These methods help with the analysis and comparison between the children's abilities to form memories, ideas, and conclusions. The findings indicate that there is no significant correlation between age and white matter structural connectivity. However, there is some correlation between white matter structural connectivity and episodic memory abilities. The correlation can specifically be seen with Hippocampus-parietal connections, meaning more mature or higher densely packed, myelinated axons between the hippocampus and parietal regions of the brain lead to better episodic memories, or at least better recollection of them.

The results also find a correlation between age and episodic memory as the 6-year-olds performed significantly better than the 4-year-olds in the memory tests, showing that the period between four and six years of age is significant for the development of episodic memory recollection, although white matter does not significantly develop during this time period.

Riggins et al. 2016 also focus on early childhood. Riggins et al. 2016 aim to understand the organizational differences in the brain and their correlation to age. The method used by the researchers to analyze the episodic memory abilities was a behavioral memory task and fMRI. The behavioral memory task was a test where the children were given toys in different rooms, and each toy was associated with a task to do, such as hugging it or drumming it. After a 1-hour delay, the scientists analyzed what the children remembered and what they did not remember. This helped the scientists understand the children's item memory abilities as well as context memory abilities. When analyzing the fMRI, the scientists focused on the hippocampus and examined its connection to other regions such as the temporal, parietal, prefrontal, and precuneus regions. In addition, both anterior and posterior regions were analyzed to see if there was any longitudinal specialization.

The results of these tests gave significant insight into episodic memory development in early childhood. The six-year-olds did much better when compared to the 4-year-olds on the behavioral memory test. This means the age correlated positively with memory scores. The fMRI shows that the hippocampal connectivity networks resemble adult memory networks, showing that the hippocampus is significant for memory, even in early childhood. Evidence of specialization also appeared, showing that different areas of the hippocampus are somewhat specialized in early childhood. A unique result appeared when analyzing hippocampal connectivity with memory regions. Six-year-olds with stronger hippocampal connections to memory regions had better episodic memory, so there was a positive relation; however, in four-year-olds, a stronger connection had a negative association with memory. An interesting discovery was connections outside of the hippocampus memory networks, positively related to memory in four-year-olds, but had a negative relation to six-year-olds.

These results show how significant the hippocampus is in early childhood and also suggest that the memory networks in early childhood are in a transitional phase, where younger children don't always use the networks that adults predominantly use, and their memory relies on

other regions as well. At six years old, the hippocampus becomes more important for memory and becomes the main region for memory-related networks. This transition also results in significant improvements in episodic memory capabilities.

Overall, early childhood is a time when the brain reorganizes its memory-related networks, becoming more specialized and starting to rely more on the hippocampus, as well as an increase in specialization and overall episodic memory abilities. This change in organizations carries through to adulthood, where similar networks can be seen, with reliance on the hippocampus being a significant development that occurs during early childhood.

The next age group covered was middle childhood. Guo et al.(2024) studied children aged 6-12 to analyze how well they performed in episodic memory tests. The researchers studied the children's episodic memory through a treasure hunt where children hid different items themselves, and after a certain time delay, the children were asked what object they hid, where they hid each object, and when they hid each object. These questions were used so the researchers could understand the WWW integrated memory. They checked whether the children could recall all 3 aspects at the same time or whether they struggled. Guo et al.(2024) ran different questioning systems, which varied in the level of support, with one version giving more support with the questioning by giving options to the children, whereas the other version gave less support and was more open-ended, requiring free recall. This allowed the researchers to analyze the strategic retrieval ability and associative binding abilities of the children.

The researchers found that the 3 aspects of memory tested improved with age being item, spatial, and temporal memory. The integration of these memories also improved with age. When support was given in the questions, the younger children benefited more, suggesting that strategic retrieval abilities begin to develop during middle childhood. The research showed that the associative binding abilities developed linearly; however, some analyses hint that associative binding abilities may develop faster around the age of 9. A hypothesized reduction in memory performance due to neural networks reorganizing at the age of 11-12 did not occur.

Bettencourt et al. 2021 studied 3 different age groups to better understand how episodic memory develops over time. The age groups are 7-9, 10-12, and young adults. The participants were given items paired with contextual information, which they were to study. During retrieval, the participants were to determine whether or not they were given a certain item, and if they were, the participant would also be expected to remember the context to go along with the item. An ERP analysis was also given to find differences when recalling old vs new items, and correct vs incorrect recollection.

The results showed that a higher age resulted in a better performance on the memory-related tests, where adults performed better than 10-12 year olds, who performed better than the 7-9 year olds. The ERP showed that the 7-9 and 10-12 age groups both had differences in the ERP readings when comparing old and new items, and the difference was in a similar fashion to the adults' readings. The difference between the children and the adults was mostly in the amplitude and the length, where the children had lower amplitudes and lasted longer. This difference in ERP readings correlates with less efficient and slower recall than adults; however, it

shows that children in middle childhood have similar memory processes when compared with adults.

Overall, the results for the experiments show that temporal context memory improves significantly during middle childhood, and ERP findings illustrate that children have a similar thought process and generate the same neural signals; however, it is just less efficient and weaker than adults. Showing even more development occurs in later stages.

Pathman et al. (2024) focused on children aged 7-11 and compared them to adults. The participants visited a museum and they photographed around 120 exhibits in around 2 hours. As the participants took the pictures themselves, the events were personalized, as well as the order. After 1-2 days later the participants looked at pairs of photographs and they were supposed to decide which photo came earlier between the two. The researchers manipulated the distance between the events. Some were short as the 2 events were taken close to each other in time, while some were longer and had larger distances between each other. The researchers also manipulated the order of the given photos, as sometimes the photos were presented in the correct order, and sometimes the photos were given in the reverse order. During this process, ERP recordings were made. The ERP recordings were made to focus on whether short or long differences in temporal order changed the ERP results.

The results showed that both children and adults have a higher accuracy when the distance between the two events was longer, and the accuracy decreased as the distance between the two events in the pictures became shorter. Another observation was that adults showed a memory asymmetry effect as they performed better when the pictures were in the correct order of events, but this effect was mainly seen when there was a small time difference between events. The ERP showed a clear difference in neural responses when comparing short and long temporal distances, and this difference occurred in both children and adults. The ERP additionally showed higher sensitivity for the order of photos given.

Overall, these results showed that temporal order for episodic memory improves from middle childhood to adulthood, and children are not as sensitive to asymmetry; affect means that sensitivity develops in later stages of life. The ERPS suggest that both groups differentiate between close together events and faraway events, showing that this differentiation in neural processes has already developed before middle childhood.

Coughlin et al. (2014) focus mainly on middle childhood with five, seven, and nine-year-olds as well as young adults. In the paper, the participants created stories about past and future events. The changes in these stories were the temporal distance, which was 1 week compared to 1 year, and the temporal direction was also changed, comparing the past and future. After this, the patients subjectively rated the story's clarity, how easy it was to remember, and their perspective when remembering or telling the story. These stories were then given to the parents of close people with the patient in order to determine the accuracy of the story or the event's plausibility.

The results showed significant improvement from the age of 5 - 9 when recalling events with contextual detail. Prospection or future stories were much less rich with contextual details

than past events, and this was true in all age groups. The future events were considered harder to create stories for, and they were usually in a 3rd person perspective. Another finding in the paper was that the quality of the episodic memory of the past would correlate with the quality of episodic memories or the prospect of the future. This means that if a person has higher-quality memories of a past event, then that person is likely to have better-quality perspectives with more contextual details.

These results signify that episodic memory recall and prospection both improve; however, prospection does not improve consistently and is consistently worse than recalling past memories. The results also suggest that children rely on episodic memory to create prospects of the future, signifying another importance of episodic memory.

Güler and Thomas (2013) focused on the age 8 - 9 year olds and 12 - 13 year olds to analyze differences between the two age groups. The researchers analyzed episodic memory by analyzing fMRI scans of children associating different images. To explain a bit more, the Children were given pairs of unrelated objects and they were told to memorize them. There were two recall phases, one was silent and the other was verbal. In the silent recall phase, the children were given one item and told to press a button if they could recall the partner in their head. With verbal recall, the children were given the same objects and told to say the paired object aloud.

The results showed that the older children performed better than the younger children. The fMRI scans, when encoding the pairs of objects, younger children used the right dorsolateral PFC and right temporal cortex more than older children, whereas older children used the left ventrolateral PFC and parietal cortex more, and these areas are associated with encoding strategies. In recall, older children used the left ventrolateral PFC and the left interior parietal cortex, and during successful recall, these areas were used more strongly. However, younger children did not show specific areas of the brain that they used significantly, suggesting a more diffuse network.

These results signify a more specialized, concentrated neural network for recall and encoding as children move through middle childhood. The PFC is important for both encoding and recall, and a higher usage of the PFC shows more effective strategies in middle childhood and the beginning of adolescence. The results also show that the maturing of PFC-parietal networks is significant for memory improvements, along with the hippocampus maturing as well.

Frick et al. studied children around the age of 8-13 to focus on middle childhood. The method used to study episodic memory was by using a free recall task. The free recall task required the children to study 20 words from 5 different semantic groups, and they then recalled them without any cues. The Adjusted Ratio of Clustering(ARC) then measured if the children's recall was organized semantically. The researchers then gave representation tasks to the children to study the children's knowledge base and semantic knowledge. Control tasks were given to measure cognitive control, which were given to measure the children's ability to apply strategies on their own. The results showed that some level of clustering does exist at the age of 8-9; however, clustering reliably improves recall at the age of 12-13. Younger children who used clustering usually did not have any benefit because of utilization deficiency, where they can not

correctly utilize strategies, although they know of them. In 8-11-year-olds, the recall was supported mainly through representation, whereas in 12-13-year-olds, the recall was supported by both representation and clustering. The ARC only predicted recall in 12-13-year-olds.

The results from the paper highlight that a shift in recall occurs as 8-11-year-olds depend on knowledge representations, while older children's recall uses strategic organization to give better results.

Overall, in middle childhood, there are significant improvements in memory overall as a whole including information such as what, when, where, and the integration of these various aspects. Through brain scans, researchers have found that the neural networks in recall are more diffused in younger children and more concentrated and specialized in older children. Younger children in the middle childhood category begin to develop strategies for recall and organization. But it is not always effective. However, children on the older side of the middle childhood reliably use strategies for organization and recall, such as semantic clustering, which helps them perform significantly better in memory tasks, along with regions of the brain maturing on their own. Younger children will benefit more from support given when recalling than older children because of this difference in strategies for recall. Older children in the middle childhood category have similar neural networks to adults. However, they are still not as efficient, and that will develop at later stages.

The final paper, by Friedman et al., covers a diverse range of age groups, being 9-10, 13-14, 20-30, and 65-85. This paper will add information for the end of middle childhood and early adolescence, as well as adults, and even seniors. The paper gave the participants a list of words, and each word was studied for later recollection. In the test phase, the participants were given a mix of words that included words they had already given or old words, as well as words not previously given or new words. For each item in the list of words, the participants were to determine whether the word was old or new and whether they recognized the item with specific details and had contextual detail to go along with it, or whether they recognized the item but did not think of it with contextual detail. This study was repeated 4 times. ERP recordings were also made to recognize familiarity versus recollection. Young adults performed the best in the study, and they also benefited the most from the repetition of the trials. Recollection also increased with repetition vs familiarity, which did not improve. The children relied mostly on recollection and did not use familiarity as frequently, while adolescents used both styles of recall. Low-performing individuals evidently had deficits in recollection, showing a significant variability in that age group.

Overall, familiarity-based recall develops at young ages around middle or even early childhood, while Recollection develops later and peaks in young adulthood, declining slowly through age.

DISCUSSION

The importance of episodic memory is evident in many domains. For example, its role in eyewitness testimony has critical implications for the justice system, while research on memory

retention informs educational practices. Episodic memory also intersects with mental health: traumatic memories can profoundly affect well-being, and understanding how such memories are encoded, recalled, and potentially modified may point toward strategies to reduce their impact (Pause et al. 2013).

Across the reviewed studies, episodic memory shows substantial improvement throughout childhood. In all age groups, older children consistently outperform younger ones on memory-related tasks, reflecting the steady development of this system across early life. During early childhood, contextual details, such as when and where an event occurred, improve markedly. The hippocampus becomes more functionally integrated within the memory network, shifting from a more diffuse to a more specialized role. White matter pathways linking the hippocampus to other regions also predict memory performance, underscoring its growing importance during this period.

In middle childhood, strategic processes that support memory recall begin to emerge. While initially inefficient, they become increasingly reliable in older children. Temporal aspects of memory also develop significantly, and episodic memory capabilities improve not only due to regional maturation but also because of large-scale brain reorganization, bringing the neural network closer to the adult pattern.

By adolescence, both familiarity and recollection are well-established and used reliably, though not yet at full adult efficiency. In young adulthood, episodic memory becomes highly efficient, with robust and differentiated patterns supporting familiarity and recollection. In older adults, recollection declines while familiarity remains relatively stable. Importantly, there is considerable variability among seniors: high- and low-performing individuals differ markedly in recollection ability.

Despite these insights, several research gaps remain. This review sought to understand developmental changes in episodic memory, yet many studies relied heavily on neural connectivity and activity measures (e.g., fMRI, ERP). Far fewer addressed how people *experience* episodic memories, such as their vividness or perspective (first-person vs. third-person). More research into these subjective aspects is needed.

Another limitation is the narrow age range of many studies. Most focused on one or two developmental stages, often leaving out transitions across the lifespan. Although challenging, studies that systematically include a broader range of age groups with comparable tasks would provide a more comprehensive picture of episodic memory development.

In addition, relatively few studies examined changes within adulthood. Young adults were often included only as a control or comparison group, and differences within adulthood were rarely analyzed in detail. This represents an important gap, given that episodic memory continues to change across the adult lifespan.

This review was also limited by its database coverage. Broader searches might have identified additional relevant studies, especially those focused on subjective aspects of memory rather than primarily neural correlates. Expanding database coverage could strengthen future reviews and help shift the focus toward how episodic memory is *perceived* as well as how it is

supported neurologically.

Future research would benefit from including wider age ranges and examining both neural and phenomenological aspects of episodic memory. Understanding not only how memory changes across the lifespan, but also how individuals subjectively experience their memories, would offer a more holistic account of episodic memory development and decline.

CONCLUSION

This paper aimed to examine how episodic memory develops over time, with a particular focus on perspective, while also addressing underlying neural changes. Episodic memory is central to human cognition and identity, shaping how individuals perceive themselves and others. A clearer understanding of both its mechanisms and phenomenology deepens not only our knowledge of memory itself but also our insight into broader aspects of human behavior and interaction.

Developmentally, episodic memory begins in early childhood with the ability to recall past events, albeit in a rudimentary form. During early and middle childhood, memory performance improves significantly as white matter connectivity strengthens and neural networks become more specialized. Strategic retrieval processes also emerge, though they are not always efficient at first. By adolescence, these strategies become increasingly reliable, and episodic memory capabilities continue to expand. In young adulthood, episodic memory reaches its peak, characterized by highly efficient and differentiated processes. With aging, some aspects of episodic memory remain stable, such as familiarity, while others, particularly recollection, begin to decline.

Overall, childhood represents a critical period for the growth of episodic memory, laying the foundation for later memory abilities. The capacities developed in these formative years not only shape lifelong memory performance but also influence identity, learning, social understanding, and mental health. Continued research into both the neural and experiential dimensions of episodic memory will further illuminate how it develops, how it changes, and how it shapes the human experience

Works Cited

- Allen, T. A., & Fortin, N. J. (2013). The evolution of episodic memory. *Proceedings of the National Academy of Sciences of the United States of America*, 110(Suppl 2), 10379–10386. <https://doi.org/10.1073/pnas.1301199110>
- Bettencourt, K. M., Everett, L. H., Chen, Y., & Pathman, T. (2021). Examining the development of memory for temporal context and its underlying neural processes using event-related potentials. *Developmental Cognitive Neuroscience*, 48, 100932. <https://doi.org/10.1016/j.dcn.2021.100932>
- Coughlin, C., Lyons, K. E., & Ghetti, S. (2014). Remembering the past to envision the future in middle childhood: Developmental linkages between prospection and episodic memory. *Cognitive Development*, 30, 96–110. <https://doi.org/10.1016/j.cogdev.2014.02.001>
- Filley, C. M., & Fields, R. D. (2016). White matter and cognition: Making the connection. *Journal of Neurophysiology*, 116(5), 2093–2104. <https://doi.org/10.1152/jn.00221.2016>
- Frick, A., Wright, H. R., Witt, A., & Taconnat, L. (2022). The underlying processes of episodic memory development: From a unique contribution of representation to the increasing use of semantic organization supported by cognitive control. *Cognitive Development*, 63, 101217. <https://doi.org/10.1016/j.cogdev.2022.101217>
- Friedman, D., De Chastelaine, M., Nessler, D., & Malcolm, B. (2010). Changes in familiarity and recollection across the lifespan: An ERP perspective. *Brain Research*, 1310, 124–141. <https://doi.org/10.1016/j.brainres.2009.11.016>
- Güler, O. E., & Thomas, K. M. (2013). Developmental differences in the neural correlates of relational encoding and recall in children: An event-related fMRI study. *Developmental Cognitive Neuroscience*, 3, 106–116. <https://doi.org/10.1016/j.dcn.2012.07.001>
- Guo, P., Carey, E., Plaisted-Grant, K., & Cheke, L. G. (2024). Episodic memory during middle childhood: What is developing? *Journal of Experimental Child Psychology*, 240, 105828. <https://doi.org/10.1016/j.jecp.2023.105828>
- Ngo, C. T., Alm, K. H., Metoki, A., Hampton, W., Riggins, T., Newcombe, N. S., & Olson, I. R. (2017). White matter structural connectivity and episodic memory in early childhood. *Developmental Cognitive Neuroscience*, 28, 41–53. <https://doi.org/10.1016/j.dcn.2017.11.001>
- Pathman, T., Shafa, A., Vogt, E. A., & Bauer, P. J. (2024). Children's and adults' memory for the order of events in a museum: A preliminary study about temporal memory and its development using photo-taking and event-related potentials. *Neuropsychologia*, 196, 108835. <https://doi.org/10.1016/j.neuropsychologia.2024.108835>
- Pause, B. M., Zlomuzica, A., Kinugawa, K., Mariani, J., Pietrowsky, R., & Dere, E. (2013). Perspectives on episodic-like and episodic memory. *Frontiers in Behavioral Neuroscience*, 7. <https://doi.org/10.3389/fnbeh.2013.00033>
- Riggins, T., Geng, F., Blankenship, S. L., & Redcay, E. (2016). Hippocampal functional connectivity and episodic memory in early childhood. *Developmental Cognitive Neuroscience*, 19, 58–69. <https://doi.org/10.1016/j.dcn.2016.02.002>

Using AI to Predict the Risk of Cervical Cancer By Bridget Parent

Abstract

We investigated how AI can be used to predict individuals most at risk for cervical cancer, and which variables are predictive of a positive diagnosis. Cancer is still one of the leading causes of death globally, for women, cervical cancer is the fourth most common cancer. However, cervical cancer is also one of the easier cancers to prevent and detect early on. The fact that there is a large population of women who are diagnosed with cervical cancer is due to the lack of resources and education surrounding the issue. Using patient data from a hospital in Venezuela, we were able to build and test various models to predict the outcome of cervical cancer diagnosis. We found that the decision tree classifier was the best performing model and easily interpretable, which made it the best choice for the machine learning model. It was simple to decipher which variables caused a positive prediction with the decision tree, and there were several steps taken to ensure the model wasn't overfitting and could perform just as well on new data. The decision tree had an accuracy score of 0.994 and an F1 score of 0.923, which was higher than the other models. The results also show that the determining variables to the positive or negative predictions of the model were a previous cancer diagnosis, early signs of cervical cancer called cervical intraepithelial neoplasia (CIN), whether or not the patient smoked, and how many years of hormonal contraceptive usage. Further research could be done to investigate the reasoning behind these variables being indicated as influential to the machine learning model.

Introduction

Artificial Intelligence (AI) is a leading technology that has the potential to drive people to greater heights due to its efficiency and adaptability. One use of AI that has been becoming more popular is integrating into the medical field. AI has been used anywhere from radiology to planning treatments. However, this paper specifically focuses on how AI can be used to help predict an individual high at risk of cervical cancer and what variables are most important in making this decision. Cervical cancer is one of the leading cancers that affects women, but it is also one of the most preventable and treatable, if caught early. Therefore, using Machine Learning as a method to emphasize the important symptoms has the potential to decrease the socioeconomic gap in diagnosis.

This problem, considering the dataset that is being used, will be a supervised classification problem since the data is labeled and the goal is to categorize the data and identify the trends with patients that are diagnosed with cervical cancer. The dataset is from the UC Irvine Machine Learning Repository, it is a collection of survey responses from a hospital in Venezuela. The dataset has already had all responses converted into correlating numerical values, making it more usable for a wider range of machine learning methods.

The goal of working with different machine learning models is to determine the model with the highest accuracy and practicality. It is best if the model is more accessible to interpret, since knowing the most important features is important.

Background

Cervical cancer is the fourth most common cancer within women, but also one of the most preventable and treatable, when caught early. Cervical cancer affects women of color disproportionately in the United States. Globally, cervical cancer is most common in developing countries and countries with less access to healthcare and reproductive resources.

Fighting cancer at any stage is imperative, and catching it early is more so important. However, the women who are most seemingly most affected by cervical cancer are also the women who have less access to medical consultations and resources to diagnose cancer before major symptoms become apparent. This is due to several factors, notably socioeconomic factors that drive disparity amongst those who have the time, money, and locational availability to get regular screenings (Amboree et al., 1549) (Montealegre). Catching cervical cancer in its earliest stages is easier when women have regular cervical screenings from clinics, however, many cannot afford the cost of the procedure. Therefore, decreasing the disparity and increasing the access to catching cervical cancer early by identifying which groups of people are most at risk and why, will benefit the women who are the most impacted.

Machine learning and Artificial Intelligence (AI) is an effective means to identify women who have a high potential to be at risk for cervical cancer. AI is typically faster at integrating through large sets of data, which would be necessary to make the identification. Additionally, once a Machine learning model is created, many different types of data can be passed through with the function of the model being the same, so others can use their information to predict whether or not they are at risk.

Another study that also approached the issue of cervical cancer with machine learning used the model to predict a diagnosis of cervical cancer (Asadi, 513-522). This study also investigated what were the major predictors. The important predictors in all the algorithms were found to comprise personal health level, marital status, social status, contraceptives, education, and caesarean deliveries.

Socioeconomic status is a proven (Amboree et al., 1549-1555) key variable in the likelihood of a cervical cancer diagnosis. This entails women who are in lower income areas and in developing countries. Having access to a tool that is free and available with an internet connection and a web browser will allow these impacted individuals to understand if they are at risk for cervical cancer.

Dataset

The dataset that was used is titled Cervical Cancer (Risk Factors) and is a collection of survey responses taken from the 'Hospital Universitario de Caracas' in Caracas, Venezuela from the UC Irvine Machine Learning Repository. There were a total of 858 responses, and therefore 858 rows of data. The dataset had 36 columns of data, one of which was a diagnosis for Cervical Cancer. Here is what the first five rows of the data looks like:

Table I. Table of the first five samples from Cervical Cancer (Risk Factors) the yes or no questions are represented in the table as 1s or 0s respectively,

Num	Age	Num sexual partners	First sexual inter-course	Num pregnancies	Smoke (y/n)	Smoke (years)	Smoke (packs/year)	Hormonal Contra-ceptives (y/n)	Hormonal Contra-ceptives (years)	IUD (y/n)	IUD (years)
0	18	4	15	1	0	0	0	0	0	0	0
1	15	1	14	1	0	0	0	0	0	0	0
2	34	1	17	1	0	0	0	0	0	0	0
3	52	5	16	4	1	37	37	1	3	0	0
4	46	3	21	4	0	0	0	1	15	0	0

STDs (y/n)	STDs (number)	STDs:condylom atosis (y/n)	STDs: cervical condylo matosis (y/n)	STDs: vaginal condylomatos is (y/n)	STDs: vulvo - perineal condylomatos is (y/n)	STDs: syphilis (y/n)	STDs: pelvic inflammato ry disease (y/n)	STDs: genital herpes (y/n)
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0

Something that is important to note is the method of survey allowed for patients to opt out of answering some of the questions if they did not feel comfortable doing so. As a result, a lot of the columns had more than 100 missing values. For the Machine Learning models to work properly, these null values needed to be replaced by another number. It was decided that using the median value of each column would work the best due to the fact there were large disparities and variation in the values, so using the median would take into account the extremities. Before the null values could be replaced, the data had to be converted into integers or floats. Even though the data appeared to be numbers visually, it was stored as objects. This caused issues when plotting the data as seen below and would cause future issues with the machine learning models.

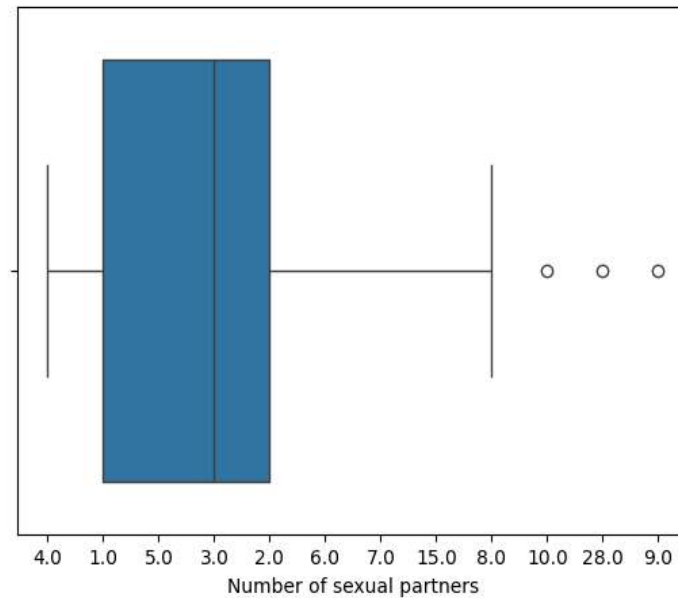


Figure I. A bar plot of the data column 'Number of Sexual Partners' before the data was transformed into integers.

The visualization of the data before it was converted into numerical data reveals why the data type could not be stored as objects. The horizontal axis of the graph did not increase equally. Additionally, the results of the two different graphs show two very different visualizations and the data stored as objects does not correctly reflect the actual distribution of values.

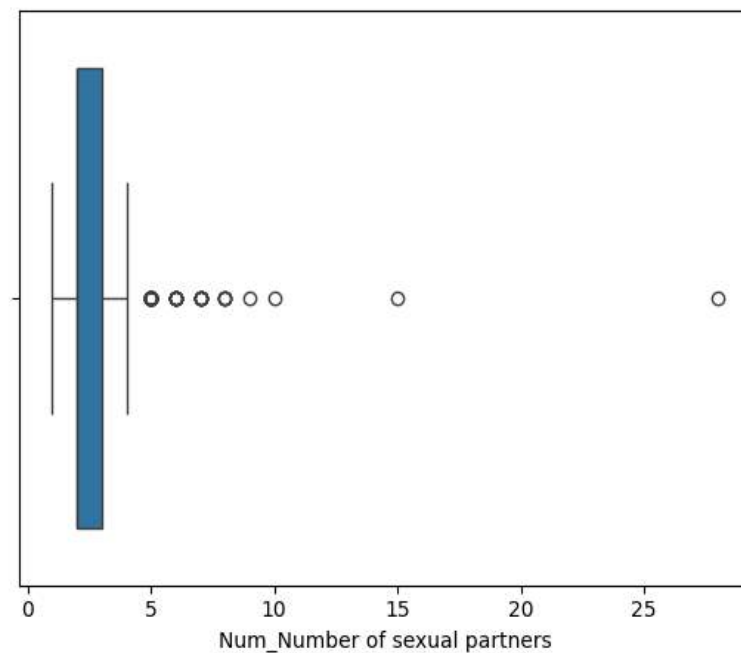


Figure II. A bar plot of the data column 'Num_Number of Sexual Partners' after the data was transformed into integers.

Additionally, for the Cervical Cancer diagnosis column, the data is unbalanced as only 24 patients were diagnosed with Cervical Cancer (at the time) versus the total 858 patients. As a result, this would likely affect the results of the Machine Learning models and the data visualization.

Method

Since the goal of the dataset was to find the relationship between one dependent variable and several independent variables, the best approach is a supervised classification model as an unsupervised model is more suited for a target value that is not set and the target is a category, not a number, so regression would not be suitable. For this dataset, a diagnosis for cervical cancer has a value of one, and no diagnosis is zero. So with the knowledge that a supervised classification model was necessary, several different types were tried.

To begin implementing the models, the dataset had to be trained and fit. The X variable was assigned to all of the columns excluding the target column of the cervical cancer diagnosis. Then, the Y variable is the target variable.

The first model that we tried is the Logistic Regression (LR) model from sklearn. LR builds off of linear regression, but for binary (categorical) outcomes of 0 or 1. The outcome variable (y) is set to equal the input variables (x_1 to x_n , where n is the number of input variables) multiplied by some regression coefficient b (b_1 to b_n) plus the intercept. This is then transformed using the logistic function, since this function has bounds between 0 and 1. Now the curve that represents the estimates resembles a “S-shaped” or sigmoidal curve. The better this curve is at distinguishing between the 0 and 1 samples, the better our logistic regression model does. This model had an accuracy score of 0.982, which is a great score but we wanted to see if we could improve our model further.

Then we tried the Ridge Classifier (RC). The RC model requires labeled data, which worked well with the dataset. The ridge classification model is designed for data with multiple target classes. The Ridge Classifier takes advantage of ridge regression which has a loss function that is determined by the squared sum of the coefficient of the model. This loss function prevents overfitting, which is an issue as when a model is overfit, it can no longer predict new data as it is just memorizing the training data. The Ridge Classifier, specifically, is a combination of the components of classification and regression to accurately fit data with multiple classes. However, the accuracy score was only slightly higher, at 0.988, and the F1 score was 0.857. Again, these scores are high, however, we still wanted to try and test if a higher score was possible.

The Random Forest Classifier (RFC) is a model that uses the combination of results of multiple decision trees to come to a singular prediction. The RFC uses decision trees, the logic is the same for the Decision Tree Classifier model, the different variables of the dataset are considered and weighed, using basic true or false analysis, the model makes a prediction based on the labeled data it is given. The RFC uses an algorithm of random vectors that makes a ‘forest’ of decision trees, the outputs of those trees are then compared. RFC trees are helpful for large sets of data since it offsets the risk of overfitting that decision trees tend to fall into (What

Is Random Forest?, n.d.). For smaller sets of data, it can over-complicate the process of making a prediction that is less interpretable. The accuracy score for the RFC was 0.988 and the F1 score was 0.833.

The final model that we tried was the Decision Tree Classifier (DTC) . DTC uses true or false to narrow down a positive or negative prediction. The model is a supervised algorithm that uses label data to create nodes and tests to decide the important variables in making an accurate prediction. DTCs are easily interpretable because of the simpler algorithm which is helpful for this paper as we want to investigate what variables are influencing a positive prediction the most. The accuracy score for the DTC was 0.994 and the F1 score was 0.923. We decided to go forward with DTC as our model because of the interpretability and high accuracy and F1 score.

Results and Discussion

Of the four models that were tested, the one that performed the best was the DT model. With the highest accuracy and F1 score of 0.994 and 0.923 respectively, it was decided that further analysis and exploration would be done with the DTC.

To first investigate the performance of the DTC beyond the accuracy score, an F1 score was used to determine the amount of times the model made a correct prediction by providing a measure of the harmonic mean of the model's precision and recall. The formula for the F1 score is:

$$F1 = \frac{2 \cdot P \cdot R}{P + R}$$

The two variables in the formula are the precision score (P) and the recall score (R). The precision score is calculated by comparing the ratio of true positives versus the sum of true positives and false positives. The closer to 1.0 the score is, the better performing the model is, and vice versa closer to 0.0. The DTC had a precision score of 1.0. When considering the desired application of this model, identifying those at risk of cervical cancer, the closer to 1.0 is important. The recall score is found by comparing the ratio of true positives against the sum of true positives and false negatives. The recall score was 0.857, which is notably lower than the precision score, but it is still considered high performance (Rey). It can be argued that this score is more important for the desired application, as it is more ideal for someone to be falsely identified as likely to have cervical cancer, than to be falsely identified to not be likely to have cancer, as it defeats the purpose of catching the cancer early on.

$$Precision = \frac{True\ Positives}{(True\ Positives + False\ Positives)}$$

$$Recall = \frac{True\ Positives}{(True\ Positives + False\ Negatives)}$$

The calculated F1 score was 0.9230. Scores above 0.9 have excellent performance as the scale of the F1 score is between 0.0 and 1.0. However, a high F1 score could also mean the model is potentially overfitting, so we continued to look at the DTC further.

We first plotted a confusion matrix, to see exactly how many false positives and false negatives there were. As shown in Figure III, the model had very few falsely predicted labels.

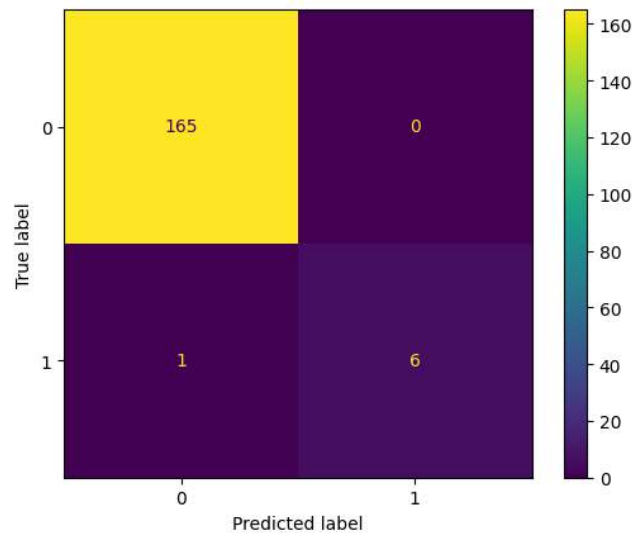


Figure III. Confusion matrix depicting true and predicted labels (1 = cervical cancer, 0 = no cervical cancer) for the decision tree model.

Then, an ROC Curve was plotted. ROC stands for receiver operating characteristic. The plot shows how well the model performs when tested. The graph clearly visualizes the relationship between how the model predicts true positives and false positives.

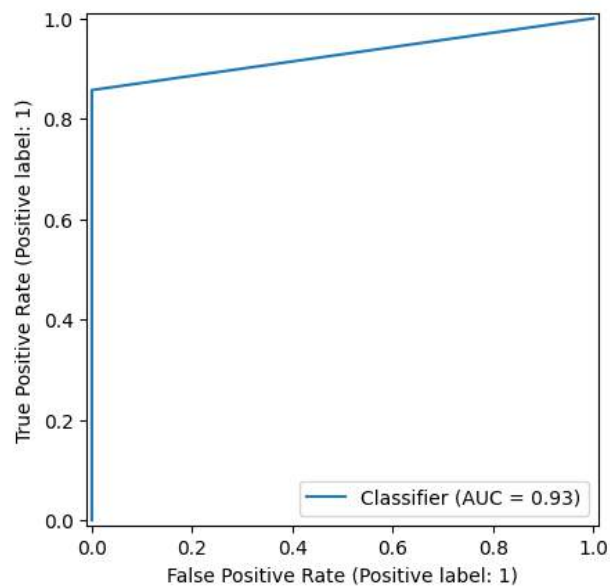


Figure IV. ROC curve used to visualize the true positive rate versus the false positive rate.

We decided to look at the different parameters of the decision tree method to see if the results could be further improved. The two parameters that were investigated were the maximum depth and the minimum weight fraction of each leaf. The depth of the Decision tree controls the longest path of a root node to a leaf node. More depth means that the tree is more complex, however the accuracy is more likely to be higher, but also leads to greater risk of overfitting.

One method to find the best parameters is RandomizedSearchCV. This model fits and scores different parameters of the machine learning method, however it does not use all possible parameters or values, unlike the GridSearchCV. The GridSearchCV makes predictions with different parameter values and then compares each prediction. The best values are then retained and returned. Both hyperparameter searches revealed that the best maximum depth was 2.0 and the best minimum weight is 0.0. Plotting the best-performing parameters of the method helped visualize what values were needed as parameters. To plot this data, a heatmap was used to see the difference in the possible values. As can be seen in Figure V, 2.0 is the best maximum depth, however, 3.0, 4.0, and 5.0, were all very close with a score of 0.99+.

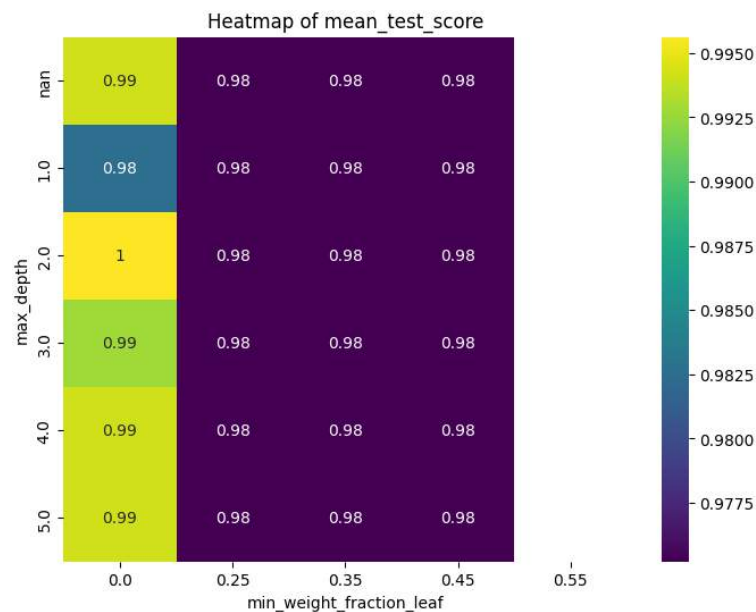


Figure V. Grid search visualization using a heatmap to compare the max_depth and min_weight_fraction_leaf parameters for the decision tree.

There was a possibility that the model was overfitting the data due to the high accuracy, precision, and recall scores. To confirm whether or not overfitting was occurring, an option to visualize how the model fits training data versus testing data, called a learning curve, was utilized. A learning curve is a graph that plots the training and the testing output. A model that is being overfit will show a loss in testing data accuracy and high loss at a turning point, this turning point is an example of overfitting (Cummings). However, the learning curve that was plotted with the results of the decision tree model did not show great evidence of overfitting,

with the testing data increasing over the samples of data and approaching the testing accuracy score of 1.0.

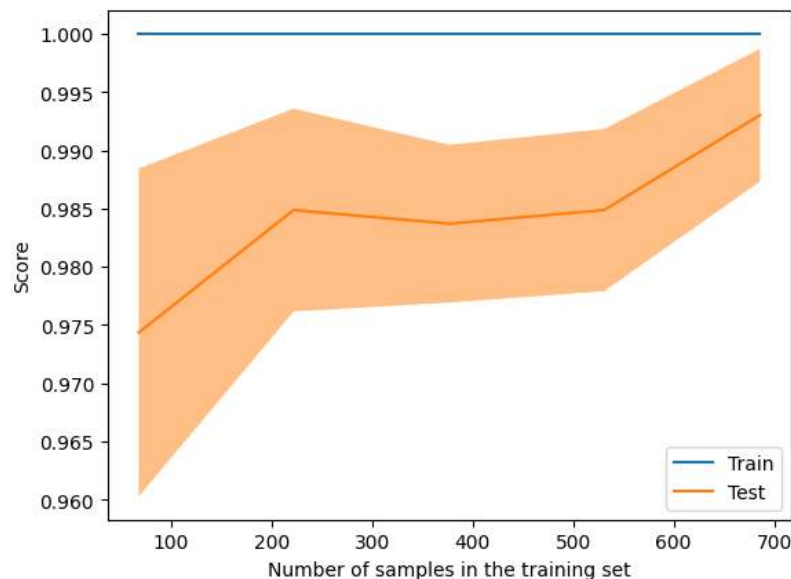


Figure VI. The learning curve of the decision tree across training and testing sets.

Having a model with high accuracy in predicting an individual at risk of cervical cancer was one of the goals of this paper, however, this paper also sought out to identify the variables that were most vital in indicating the result of the decision tree's prediction. Knowing what variables are weighed most heavily in the prediction can be useful for those who are at risk to understand why they are at risk and what further symptoms to look out for. The method that was used to go about identifying these variables is the Mean Decrease in Impurity (MDI). MDI is the sum of the amount of times a feature is split in proportion to the total number of samples that are split.

The result of visualizing the data reveals that the variables that contributed most to the predictions were years of hormonal contraceptives use, if the patient smoked, diagnosis of cervical intraepithelial neoplasia (CIN), and a previous diagnosis of cervical cancer or another kind of cancer. The most important variable was a previous cancer diagnosis with a diagnosis of CIN following closely behind. This is a logical result as CIN is the earliest indicator of cervical cancer, and is what is screened for during pap smears. Smoking has been proven to increase risk of cancer, so it is also logical that it is of high importance. However, what was interesting to us was the correlation between positive cervical cancer diagnosis and the number of years hormonal contraceptives are used. Further investigation is needed to evaluate why the number of years of hormonal contraceptive use is identified as an important feature. This is a potential avenue for future research.

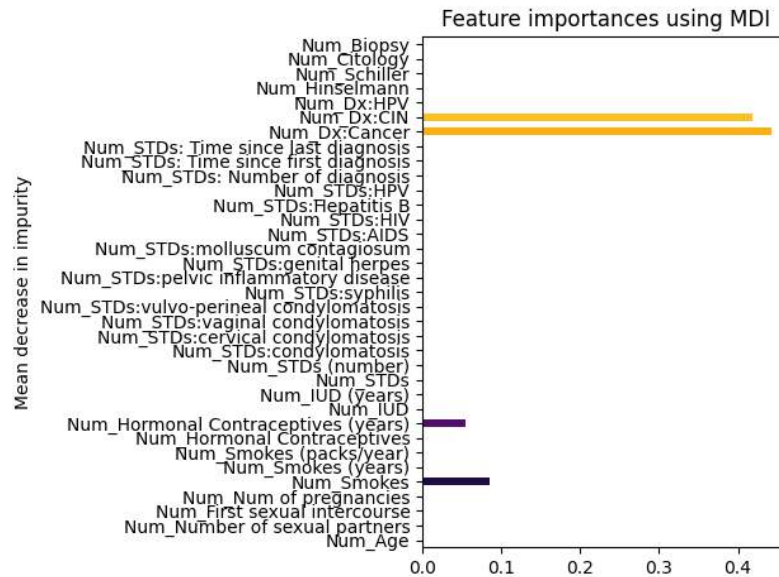


Figure VII. Graph that visualizes the importance of the columns in the decision tree model.

Knowing the individual importances of the variables was helpful, however, to investigate how the decision tree is weighing those variables we used Sklearn to plot the process of the decision tree. The plot of the tree not only indicated variables were considered for a positive prediction, but also was values of those variables. This was extremely helpful in identifying how a person may be at risk for cervical cancer. For example, the first variable, and most important for the decision tree was a previous cancer diagnosis with a value greater than 0,5. This makes sense as that data column only had two possible values, 0 indicating no previous diagnosis, and 1 indicating a diagnosis, if the sample did have a previous diagnosis, then the model looked at whether or not the sample was diagnosed with CIN, again looking if the value is greater than 0.5. If they were, the model predicted a positive for a diagnosis for cervical cancer, if not, then the model predicted a negative diagnosis. If the sample did not have a previous cancer diagnosis, the model then analyzed if the sample smoked. If they did, then the model looked at the number of years that hormonal contraceptives were used. If the value was greater than or equal to 12.5, then the model predicted a positive diagnosis, if it was less than 12.5, it predicted a false diagnosis. Therefore, the model has indicated that an individual who has used hormonal contraceptives for at least 12.5 years is more at risk than those who aren't.

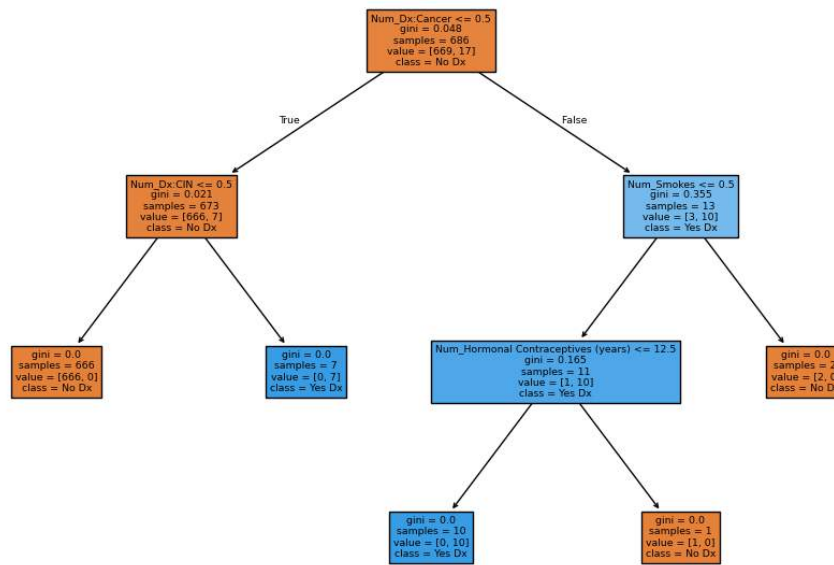


Figure VII. Visualization of the decision tree model with the Yes Dx standing for a positive cervical cancer diagnosis and No Dx standing for a negative diagnosis.

One interesting result from the feature importances, was the fact that Hormonal Contraceptive use over 12.5 years caused a positive prediction if the sample also smoked. From the American Cancer Society, taking oral contraceptives for more than five years can lead to a slight risk of cancer, and the risk rises with the longer the use of contraceptives. UCLA health also warns that, “If you use birth control pills, your probability of developing cervical cancer increases slightly. Your risk continues to grow the longer you take oral contraceptive pills,” (Sridhar). However, with typical and regular screenings, there is no need for increased concern. This makes sense as to why this feature has some importance for the decision tree, but is the feature with the least importance, compared to the other three.

Conclusion

After testing five different models, the decision tree model is the best fitting model test for the purpose of identifying those at higher risk of cervical cancer. The decision tree model had the best accuracy and interpretability that fit with the goal of identifying those at higher risk of cervical cancer and the variables that are the most important to that prediction. Several different factors of the decision tree were looked at, from the parameters, to how the tree made its predictions. The most significant result was plotting the importance of the variables, the plot highlighted what variables were likely to lead to a positive cervical cancer prediction. These variables were a previous cancer diagnosis of another kind, if the patient smoked, if they showed early signs of cervical cancer, and if they used hormonal contraceptives for more than 12.5 years. All of these features seem likely to hold high importance due to the fact there is strong

correlation between a cancer diagnosis and these attributes, except the amount of years of hormonal contraceptives. However, researching this further revealed that hormonal contraceptives can increase the risk, very slightly, and the risk increases the more years they are used. This is an important determination from the decision tree.

The goal with using machine learning to predict a diagnosis of cervical cancer is to alleviate the disparity in which women are diagnosed with terminal cervical cancer. Since the gap between socioeconomic status has a major effect on development of cervical cancer, using machine learning to make those at risk aware they are at risk can help women catch their cervical cancer early. This research can motivate the groups most at risk to get more screenings. More work needs to be done on translating these research findings into clinical practice, especially in rural areas or areas lacking access to medical professionals. Another potential area of innovation is to create an easily accessible website or application for people to assess their individual risk for cervical cancer and motivate them to get screenings more regularly.

Acknowledgements

This research was possible due to inSpirit AI which provided the resources and Neha Anand Simha who guided the research. She was pivotal in the development of the code and the paper.

Additionally, Laura Stamp, the Lewis Palmer High School GT Facilitator and my teacher gave me the time and the tools to complete this research.

Works Cited

- Amboree, T. L., Damgacioglu, H., Sonawane, K., Adsul, P., Montealegre, J. R., & Deshmukh, A. (2024). Recent trends in cervical cancer incidence, stage at diagnosis, and mortality according to county-level income in the United States, 2000–2019. *International Journal of Cancer*, 154(9), 1549-1555. <https://doi.org/10.1002/ijc.34860>
- Asadi, F. (2020). Supervised Algorithms of Machine Learning for the Prediction of Cervical Cancer. *J Biomed Phys Eng*, 10(4), 513-522. <https://pmc.ncbi.nlm.nih.gov/articles/PMC7416093/pdf/JBPE-10-513.pdf>
- Cummings, J. (2025, February 28). What Is a Learning Curve in Machine Learning? Baeldung. Retrieved March 11, 2025, from <https://www.baeldung.com/cs/learning-curve-ml>
- Jain, S. (2024, October 17). F1 Score in Machine Learning. GeeksforGeeks. Retrieved March 10, 2025, from <https://www.geeksforgeeks.org/f1-score-in-machine-learning/>
- Learning Curve Examples. (2020, March 24). Ryan Wingate. Retrieved March 10, 2025, from <https://ryanwingate.com/intro-to-machine-learning/supervised/learning-curves-examples/>
- Lee, C. (2017, October 28). Feature Importance Measures for Tree Models — Part I. Medium. Retrieved March 11, 2025, from <https://medium.com/the-artificial-impostor/feature-importance-measures-for-tree-models-part-i-47f187c1a2c3>
- Montealegre, J. (2024, January 25). Cervical cancer rates rising in low-income U.S. counties. MD Anderson Cancer Center. Retrieved January 14, 2025, from <https://www.mdanderson.org/newsroom/cervical-cancer-rates-rising-in-low-income-u-s--counties.h00-159694389.html>
- Parekh, B. (2025, January 16). Cervical Cancer & Health Equity: Bridging The Gap In Access To Care. BW Healthcare World. Retrieved January 21, 2025, from <https://www.bwhealthcareworld.com/article/cervical-cancer-health-equity-bridging-the-gap-in-access-to-care-545097>
- Rey, M. (2023, June 22). Recall and Precision: how is the performance of a prediction model measured? MadKudu. Retrieved March 10, 2025, from <https://support.madkudu.com/hc/en-us/articles/4407471867405-Recall-and-Precision-how-is-the-performance-of-a-prediction-model-measured>
- Sridhar, A. (2021, February 18). Understanding the link between birth control pills and potential reduced risk of certain cancers. UCLA Health. <https://www.uclahealth.org/news/article/understanding-the-link-between-birth-control-pills-and-cancer-risk>
- What Is Random Forest? (n.d.). IBM. Retrieved March 11, 2025, from <https://www.ibm.com/think/topics/random-forest>

Electromagnetic Interference (EMI) from Consumer Devices and Its Impact on Implantable Pacemakers with AI for Signal Classification

By Gia Bhatia

Abstract

Electromagnetic waves from everyday devices—including smartphones, WiFi routers, microwaves, and electric vehicles—pose potential risks to individuals with medical devices such as pacemakers. Electromagnetic Interference (EMI) can induce mimicking signals, mode switching, signal disruption, and temporary or permanent malfunctions, potentially compromising patient safety and delaying clinical intervention. This study investigates the interactions between EMI emitted by common electronic devices and pacemaker operation using a Ventricular Paced, Ventricular Sensed, Inhibited (VVI) pacemaker prototype on a breadboard. Data collected from controlled exposure experiments was analyzed with AI-based signal classification algorithms to identify interference patterns. Results indicate that devices emitting high-frequency electromagnetic waves in close proximity to pacemakers cause the most significant interference. Moreover, AI analysis enhances detection accuracy, enabling proactive mitigation strategies and the potential for real-time device alerts. These findings underscore the importance of integrating advanced signal monitoring and AI-assisted analysis into pacemaker safety protocols to improve patient protection in environments with pervasive electromagnetic exposure.

Introduction

Three million. That is the approximate number of people in the United States who rely on pacemakers to maintain normal heart rhythms. In today's technology-driven world, electronic devices—from smartphones and WiFi routers to electric vehicles—have become ubiquitous, offering convenience and connectivity like never before. Yet, alongside these benefits comes a less visible risk: electromagnetic interference (EMI). Research has linked EMI to biological effects such as headaches, dizziness, and nausea. In some cases, EMI can also cause more serious health concerns. For individuals with implantable medical devices, even minor EMI exposure can disrupt device function, causing unintended mode switching, false sensing, or temporary malfunctions. As technology becomes increasingly pervasive in modern society, the likelihood of such interference rises, highlighting a critical area of concern for both patients and healthcare providers. This study examines the effects of EMI on a VVI pacemaker simulation, leveraging artificial intelligence for signal classification to detect and analyze interference patterns. The ultimate aim is to improve patient safety and device reliability in an increasingly connected world.

History

The modern pacemaker is a small device implanted in the chest that delivers electrical pulses to regulate slow or irregular heart rhythms. Pacemakers are often confused with

implantable cardioverter-defibrillators (ICDs), which are also chest implants but function primarily to monitor dangerously high heart rates and deliver corrective shocks through defibrillation. Today, approximately three million Americans live with pacemakers, and about 200,000 new devices are implanted each year.

The history of pacemakers dates back to 1958, when Swedish cardiac surgeon Åke Senning successfully implanted the first fully internalized device. Although that early pacemaker failed within hours, it established the foundation for modern cardiac pacing technology. Since then, designs have evolved dramatically, with devices now lasting 10–12 years, depending on the patient's reliance. Pacemakers also come in several types, including VVI, VVO (Ventricular pacing), and DDD/DDDR (Dual-chamber, Dual-chamber sensing, and Dual-chamber response), each designed to meet specific cardiac needs. This study focuses on the VVI pacemaker, the most basic configuration, as it provides a useful platform for testing electromagnetic interference.

EMI is defined as the disruption of pacemaker function by external electromagnetic signals. EMI can cause false sensing, unintended mode switching, or pacing inhibition. While modern devices are equipped with shielding and filtering mechanisms that make such interference rare, the growing presence of high-frequency electromagnetic fields from consumer electronics raises renewed concerns.

One illustrative case involved an 85-year-old woman with a dual-chamber pacemaker who reported persistent fatigue. Her device, programmed for DDDR pacing, was repeatedly found to have reset into VVI mode, significantly reducing its responsiveness. After ruling out other causes, clinicians discovered that the patient had recently begun using an electric warming blanket during sleep. Once she discontinued the blanket, the pacemaker returned to normal function, and her symptoms resolved.

Although rare, cases like these underscore how the rapid integration of new technologies into daily life can create unforeseen interactions with medical devices. Continued research into EMI effects is therefore critical to ensure pacemaker safety in a technology-saturated environment.

Methods and Circuit Design: Overall Approach

Designing a simple VVI pacemaker simulation on a breadboard required multiple iterations and refinements throughout the project. The final system consisted of four main components: (1) an astable 555 timer circuit; (2) a potentiometer-controlled resistor to adjust the simulated heart rate; (3) an Arduino Uno serving as the pacemaker's processing unit to monitor rhythm irregularities; (4) a monostable 555 timer circuit to generate pacing pulses when triggered by the Arduino.

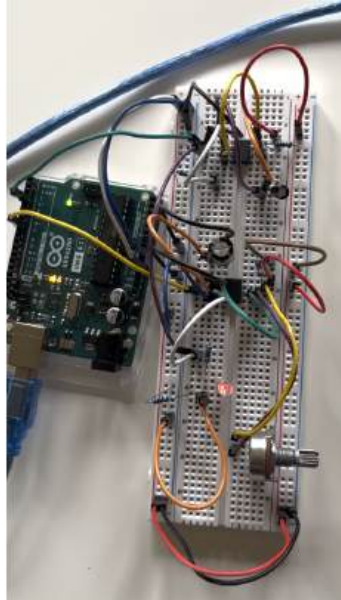


Figure 1: VVI Circuit

To assess EMI susceptibility, the circuit was tested under both shielded and unshielded conditions. Shielding was introduced by placing a layer of aluminum foil over the circuit or between the EMI-emitting object and the pacemaker circuit. In addition, an open probe connected to the Arduino Uno functioned as an EMI detector, enabling measurement of interactions between external electromagnetic fields and the pacemaker simulation.

The table below shows the parts of a real VVI pacemaker mapped to the simulated portions of the VVI pacemaker, along with a brief description of the component's overall function in the pacemaker.

Table 1: Real to Simulated VVI Mapping

Real VVI Pacemaker	Simulated Arduino Pacemaker
Pulse Generator (sends electrical signals to the heart)	555 monostable timer (fixed-length pulse, like one used to pace the heart's ventricle, generated electronically)
Heart sensing circuitry/decision-making logic (detects natural heartbeats)	Arduino Uno + code (monitors ventricular activity, decides when to send pulse) → plays the role of the pacemaker's control unit (microcontroller)
Leads, wires & electrodes (sends/receives heart signals from tissue)	Wires connecting Arduino/555s/LEDs (wires/breadboard connections mimic how signals travel out of the pacemaker to the heart; probe attached to Arduino Uno to simulate a lead)
Output to heart (delivers pacing signals to the heart muscle)	LED (visualizes pacing output, "stimulating ventricle")

Heart electrical activity (Ventricular depolarization → causes the ventricles of the lower heart to contract via electrical impulses)	555 astable timer (rhythmic electrical pulses → Electrocardiogram (ECG) signal, simulating beating heart)
Rate sensor/patient input	Potentiometer to vary heartbeats (simulates a person at rest vs exercise, → tests pacemaker logic)
telemetry/diagnostic/data output	Arduino Uno (outputs pulse signals & timing data for EMI analysis)

Heartbeat Simulation

The first step in developing the pacemaker system was to establish a simulated heartbeat. For this purpose, a 555 timer was configured in astable mode, a non-stop oscillator that creates a continuous, adjustable square wave to simulate the natural heartbeat. By experimenting with different resistor-capacitor (RC) configurations, a stable output was achieved at approximately 60 beats per minute (bpm), corresponding to a typical resting human heart rate.

The circuit consisted of two resistors in series paired with a capacitor, forming the RC timing network that governed the output pulse frequency. The resulting periodic square wave served as the intrinsic electrical activity of the heart.

Figure 2 below illustrates the astable circuit on the breadboard, currently being powered by a 9-volt battery. The resistor values initially used were 10 k ohms as R1 and 100 k ohms as R2. The capacitor value was 10 microfarads. The LED in Figure 2 that is lit serves as the visual signal of the heart rate and will later also be used as the visual output for the pacemaker's pulses. The components in Figure 2 consist of two resistors, a capacitor, and one red LED, all powered by a 9V battery.

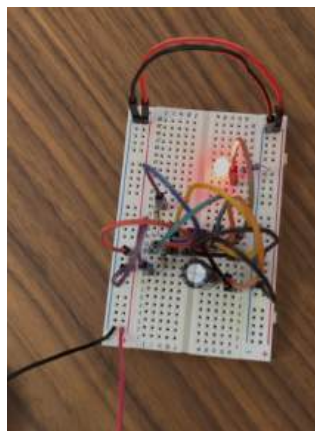


Figure 2: Astable 555 timer

Rate Variation

After configuring the heartbeat simulation, R2 was switched out on the astable 555 timer for a 100k ohm potentiometer to tune the oscillation frequency and allow a change in the bpm of

the heartbeat. The potentiometer functions by limiting the amount of current able to pass through at a given moment. The potentiometer tuned for higher resistance will result in a slower heart rate, while a lower resistance will result in a faster heart rate. The 100k ohm potentiometer's maximum resistance value is 100k ohms. This feature was implemented to allow testing of the pacemaker's ability to respond to bradycardia (slow heart rate), and tachycardia (fast heart rate).

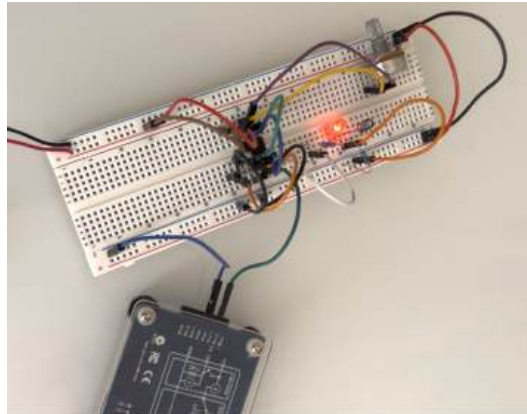


Figure 3: Astable 555 timer with potentiometer

Monitoring and Pacing

Once the heartbeat and a method to test the functionality of the pacemaker logic were established, the pacemaker logic needed to be implemented. The output of the 555 astable timer was connected to an Arduino Uno input pin. The Arduino was programmed to monitor the time between successive pulses, compare this to a preset threshold of 80 bpm, detect missed beats or prolonged pauses, and trigger an external pulse through the monostable 555 timer if any anomalies are detected. The functionalities listed above are congruent with the VVI pacemaker logic, which monitors the ventricles and delivers pacing when the heart rate drops below the pacemaker's programmed rate.

The monostable 555 timer was set up such that the Arduino Uno's output pin was fed into the monostable 555 timer's input pin, and the output pin of the monostable 555 timer connected to the LED. Figure 4 below depicts the astable 555 timer (circled in yellow) next to the monostable 555 timer. The monostable setup ensured each triggered pulse had a fixed, stable width, independent of trigger duration. The combination of the monostable configuration and the LED visual output effectively simulated the pacing function of a clinical pacemaker. The loose wires will be attached to an Arduino Uno.

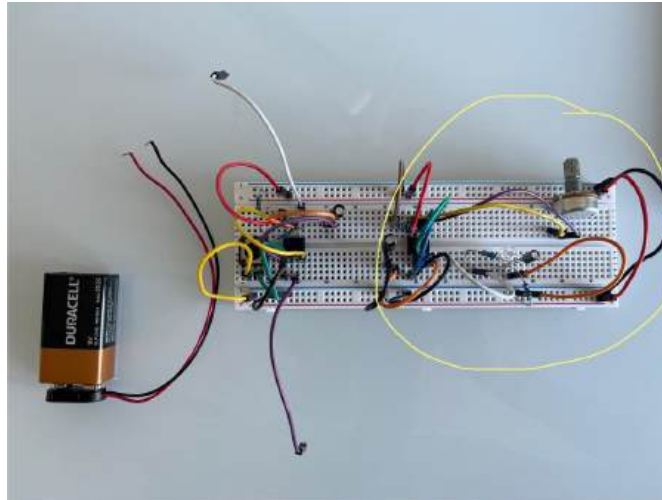


Figure 4: Astable (right) and monostable (left) 555 timers

Coding VVI Logic

The Arduino Uno was coded using the Arduino Cloud IDE. The `pinIn` and `pinOut` functions were defined on the Arduino Uno and were connected to the astable and monostable 555 timers, respectively. A constant was utilized to keep track of the interval threshold of 80 bpm in milliseconds — about 750 milliseconds — and declared as a global variable that keeps track of the last beat time in milliseconds.

In the setup method, the `pinMode` and `digitalWrite` were utilized to configure the Arduino Uno's idle state, start the Serial Monitor, and delay it to allow the circuit to establish a rhythmic heartbeat before counting the time between beats. An anomaly that occurred was having the digital write respond to the opposite call, meaning if the output of the 555 monostable timer is to be off until a specific call, the digital write would need to be HIGH for the output pin, not LOW. This syntax continued in the loop method, where the program would first check whether a beat is detected, and inhibit pacing thereof. Next, the code checks whether too much time has passed without a beat by subtracting the total amount of milliseconds the code has been running from the recorded milliseconds of the last beat time, and comparing the result to the interval threshold. If too much time has passed, the digital write function is used to set the `pinOut` LOW, then HIGH, and ends by resetting the last beat time to the number of milliseconds the circuit is currently at. Finally, the code concludes by setting up the serial plotter display and finishes with a delay of 200 milliseconds for the pacemaker to debounce any of its own pacing. You can find the code here: https://github.com/giabhatia/VVI_Pacemaker

Shielding & EMI Detection Setup

To replicate shielding on real pacemakers, aluminum foil was wrapped around portions of the circuit while testing EMI. Tests both with and without shielding were conducted, and the results of both were compared to roughly determine the efficiency of EMI shielding against EMI.

Figure 5 below shows a makeshift probe (circled in yellow), placed on the Arduino Uno to detect external EMI around the circuit and receive visual input on the serial plotter. This feature served to detect variations in waveform amplitude and frequency when exposed to different sources. The probe simulates a lead on a pacemaker, which is also susceptible to EMI, displayed in the image on the right.

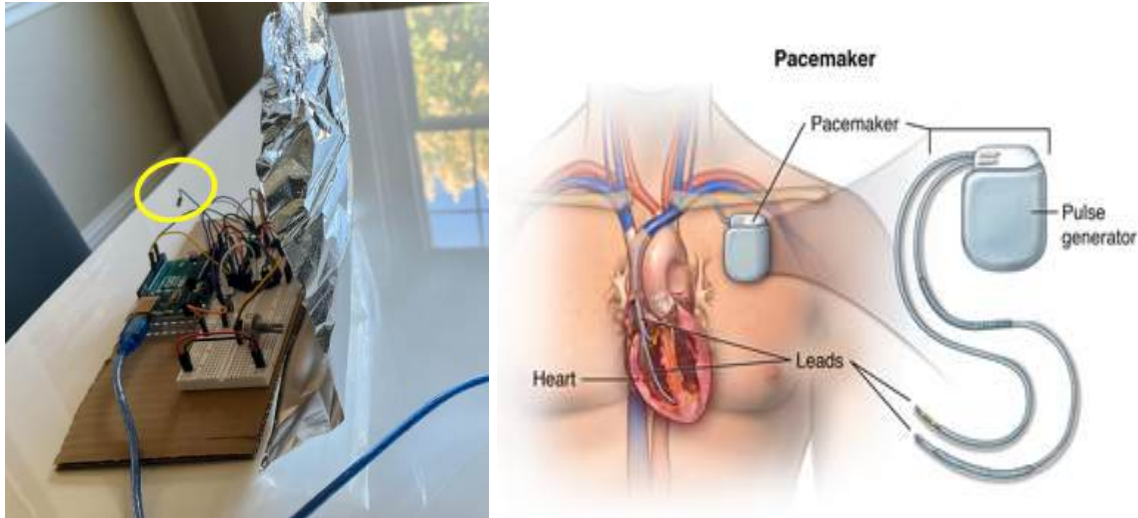


Image of completed circuit with the probe circled in yellow next to an image of a pacemaker with leads taken from Johns Hopkins Medicine.

Figure 5: Circuit with probe (left) and pacemaker with leads (right)

The VVI pacemaker simulation circuit was brought near EM emitting objects, where data points were recorded from the following:

- Apple MacBook Air
- Apple iPhone Mini 13
- GE Microwave
- High EMI desk setup: sit-stand desk, table lap, and study timer
- AirPods 1st Generation
- WiFi Router
- Tesla Car Model Y

Data Collection & Trials

Waveforms from the pacemaker circuit were observed and logged via the Arduino Uno IDE serial plotter. In each trial, the circuit was exposed to an EMI source up to a maximum distance of 4 meters away from the circuit. Simultaneously, screen recordings were taken of the serial plotter for analysis, as well as utilizing the serial monitor, which outputted numeric data to be later fed into artificial intelligence (AI) models for data analysis. Interference, such as missed beats, extra pacing pulses, or signal distortion, was analyzed in the following trials. The Arduino Uno does not measure in units on the y-axis, but utilizes milliseconds on the x-axis.

The laptop was connected to the Arduino Uno, providing both power and ground connections via USB. Upon connecting the probe to the Arduino Uno, the laptop's EMI was immediately detected and picked up significant interference. The graphs below show the difference in interference picked up on the serial plotter when the probe was attached versus when it was not attached to the Arduino Uno. The single spikes are the pulses the VVI pacemaker is sending to the LED, while the approximately steady pulse is the astable 555 timer's pulse to the LED, turning it on.

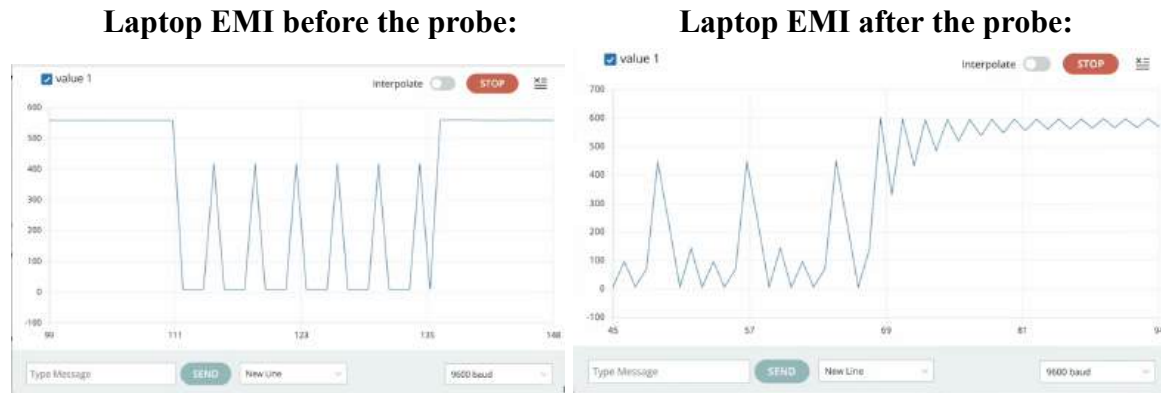


Figure 6

The spikes are caused by the Arduino Uno being electrically grounded to the laptop, as the USB cable provides both power and ground to the Arduino. The interference shown is expected as the laptop's ground can be noisy due to power supply switching, USB bus chatter, and processor activity. Therefore, even when the shielding was applied, the same level of interference appeared on the serial plotter, as the EMI is not just being radiated by the laptop, it is also injected through the USB cable. As a result, when the data was analyzed, it was based on the interference of the laptop being present at all times.

iPhone Calling EMI (Information about the Shielding vs Non-Shielding Oscillation Data)

The iPhone EMI was recorded by placing an iPhone about two inches from the probe with its speakers facing the probe. A call was placed on the iPhone set on speaker mode to see if the vibrations could cause interference.



Figure 7: iPhone EMI Testing

The serial plotter depicted both EMI signs when tested with shielding and not shielding. A key finding includes there being a spike during both shielding and non-shielding trials between the monostable and astable timer transition. Comparing the trials both with and without shielding, there appears to be roughly the same number of spikes. This shows that the shielding did not fully block EMI from the surroundings. Upon close inspection of the serial plotter, there appears to be a little less interference concerning the astable 555 timer portion, as there is less interference seen in the stable line. The spike would appear in other EMI tests as well — some occurred during all the transitions from the monostable 555 to the astable 555 timer, while others did not occur as frequently. This spike could have also come from hardware issues. However, since the spike was not consistent and did not occur for each test, it is reasonable to consider this a sign of EMI.

iPhone EMI no shielding with the spike:

iPhone EMI no shielding without the spike:



Figure 8

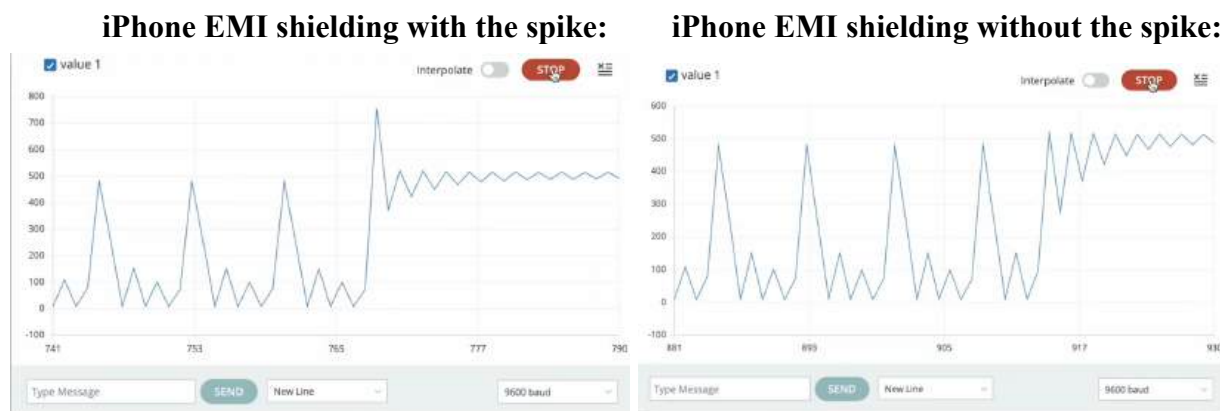


Figure 9

The AirPods trial resulted in very little interference, as shown by the serial plotter, other than the occasional spike. Minimum interference by the AirPods is expected, as they emit high-frequency EM waves, and the probe may not be that sensitive to higher-frequency EMI. However, upon comparing the serial plotter's data concerning shielding and no shielding, the spike appeared much more when the circuit was shielded than when it was not. This led to some concern as the shielding is meant to reduce EMI on the circuit. For this reason, the verification of the hardware components in the circuit was necessary. Noting that the hardware was in order, human error could have occurred, such as the aluminum foil resting on the probe during the shielding tests, thus causing the sensitive circuit to spike upon contact.

Upon close inspection of the microwave used for testing, there appeared to be a Faraday cage encapsulating the internal box of the microwave. Faraday cages are often used to replicate the encapsulation of pacemakers as well. Therefore, when conducting the microwave trials, no EMI effects were expected to be seen, which was the case for the no-shielding trial, but a spike was seen for the shielding trial. Yet again, this brought up the question of whether this was a hardware issue or another human error. Seeing no issues with the hardware, this could have occurred due to improper shielding placements. Possible scenarios could be due to not covering the entire circuit with aluminum foil, or EMI from nearby EM-emitting objects, such as the fridge and oven. That being said, not all microwaves have a built-in Faraday cage to contain EM waves, as some older or worn-out microwaves could experience some leakage with their Faraday cage, thus potentially posing a threat to EMI.

WiFi EMI (Information about the Shielding vs Non-Shielding Oscillation Data)

The WiFi router's serial plotter output was as expected. When the circuit was not shielded, there was distinct interference compared to when shielding was utilized—mainly seen in the astable part of the visual output. This is accounted for, as the WiFi router radiates substantial EMI to surrounding objects, in this case, seemingly affecting the amount of noise the Arduino Uno receives from the astable 555 timer.

WiFi EMI no shielding astable 555 configuration: WiFi EMI shielding astable 555 configuration:

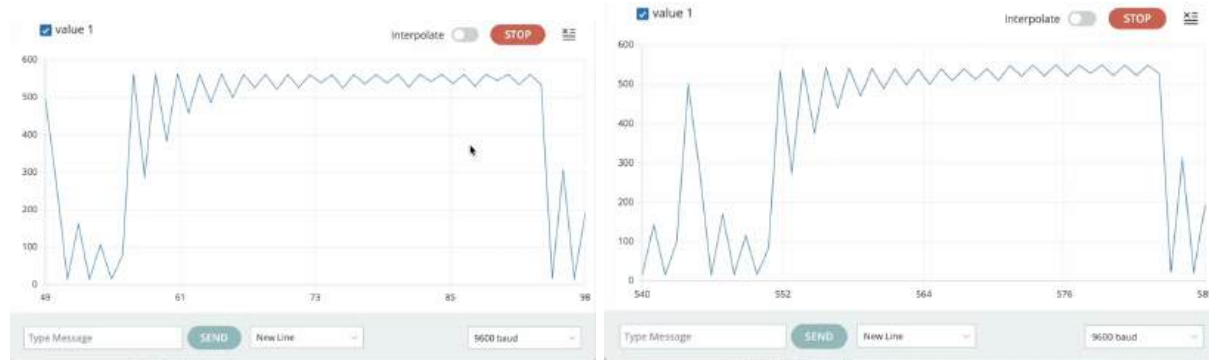


Figure 10

Tesla Car Seat EMI (Information about the Shielding vs Non-Shielding Oscillation Data)

Both front and back-seat EMI were tested in the Tesla Car, during which the spike was observed again. The Tesla car was parked in the garage when the tests were conducted, not charging, the AC was on at 69 degrees, and had the internal car lights on to illuminate the circuit. Furthermore, the circuit was placed in the middle console when tested in the front, and the middle seat when tested in the back. Based on the Model Y the circuit was tested on, the battery of the car is in a “skateboard” platform, meaning the battery was distributed roughly equally throughout the car. There are known to be batteries in the frunk of the car, and overall more electrical components towards the front of the car, such as the control panel touch-screen.

Tesla Car backseat no shielding:

Tesla Car backseat shielding:



Figure 11

The “no shielding” image shows how the spike is present, and the output pulses are higher than expected, which is greater than 500. The same follows for “with shielding.”. The spike was constantly seen during the transition of the Arduino output, the monostable 555 timer, to the Arduino input, the astable 555 timer. During the trials with shielding, the spike was seen less consistently, though still present.

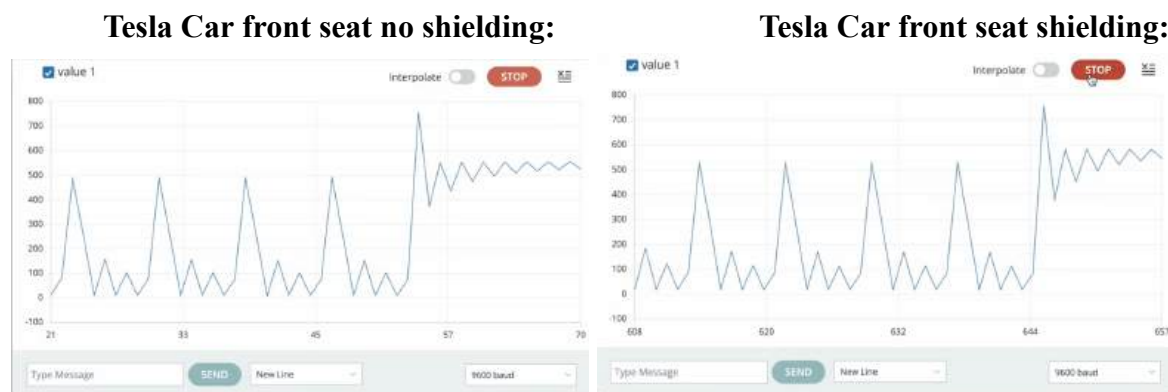


Figure 12

The images above show that the shielding trials resulted in a higher output pulse (greater than 500) compared to the non-shielding trials, and the spike that occurs both with and without shielding. The input astable configuration appears relatively undisturbed compared to the output Arduino pulses and the monostable pulses, as the period when the astable timer was measured was quite stable.

Surprisingly, the shielding trials for both front and back seats showed a much higher record of EMI. This can be explained, as the shielding could have acted as a resonant cavity due to improper grounding. It may be hypothesized that, rather than blocking EMI, the shielding causes the waves to be trapped and reflected within it, thereby increasing the total electromagnetic field strength.

These trials highlight that shielding is a necessary component for pacemakers to block electromagnetic interference, but the Tesla trials demonstrate how the shielding must be done currently, as an incorrect implementation could result in increased EMI.

The High EMI desk setup consisted of the circuit placed in the center of the sit-stand desk, occasionally moving up and down, a table lamp to its side in the lowest mode, and a stopwatch, all operating simultaneously. The spike was prominent in both shielding and non-shielding, continuously showing up between the transition from the Arduino's output to the astable 555 timer's output.

High EMI Desk no shielding:

High EMI Desk shielding:



Figure 13

Once these trials were conducted, the data from the serial monitor were input into AI models for data processing.

Artificial Intelligence Implementation

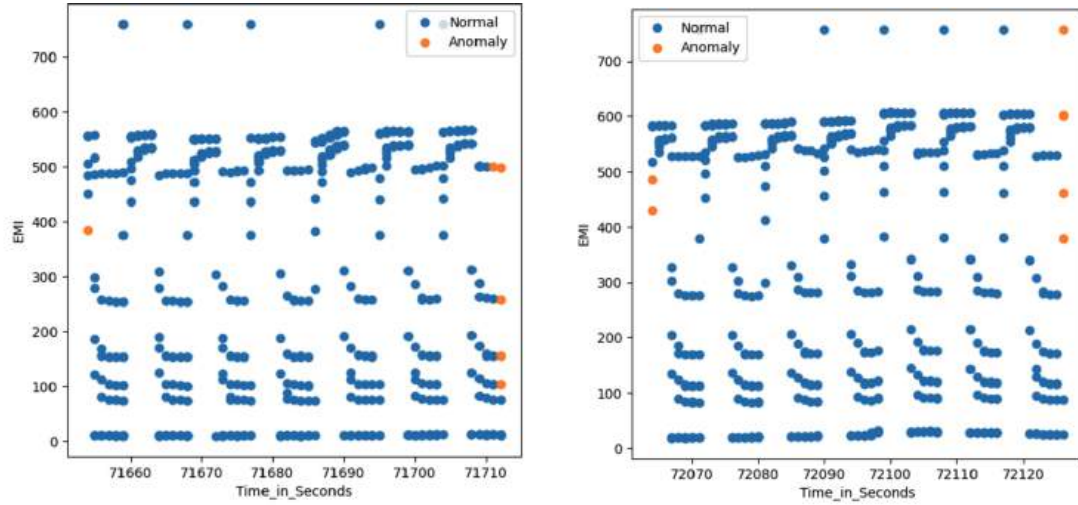
The AI models used for data analysis include Isolation Forest, Random Forest, and Autoencoder Anomaly Detection Neural Networks. Due to the high volume of data from the serial monitor and time constraints, only the Tesla Front Seat, WiFi, and iPhone inputs were sorted into data sets. All the EMI data sets from the serial plotter are available in the GitHub repository.

Creating the data sets included extracting the EMI and time in milliseconds and converting them to seconds, from the serial monitor. The 5-number summary –consisting of the minimum, quartile one, median, quartile two, and maximum– was taken for the EMI data points, and the frequency of the occurrences of EMI data points within these categories was noted. Furthermore, the low and high outdoor temperatures, as well as the garage temperatures, were recorded. These values were held constant in their respective category (low or high). The same was true for the outdoor dewpoint and the garage dewpoint. Dewpoint and temperature were considered to see if these values impacted the pacemaker’s functionality. Finally, the blood pH of a healthy human (values ranging between 7.35 to 7.45) and the elevation above sea level were recorded from low to high and high to low, respectively, in two columns. These last two values were considered as the blood pH could mimic internal human body conditions within the heart and provide a chemical element, while the elevation above sea level could affect breathing and blood pressure. In essence, these conditions were used to determine how this pacemaker would function in the human body.

Since the blood pH and elevation above sea level had data points that varied over time, these were analyzed in graphs against EMI. An estimate of 100 trees, 0.01 expected proportion of anomalies, was used across all models, consisting of 500+ rows of data. The variance between the number of anomalies pointed out shows why multiple models run the same data– to avoid any bias.

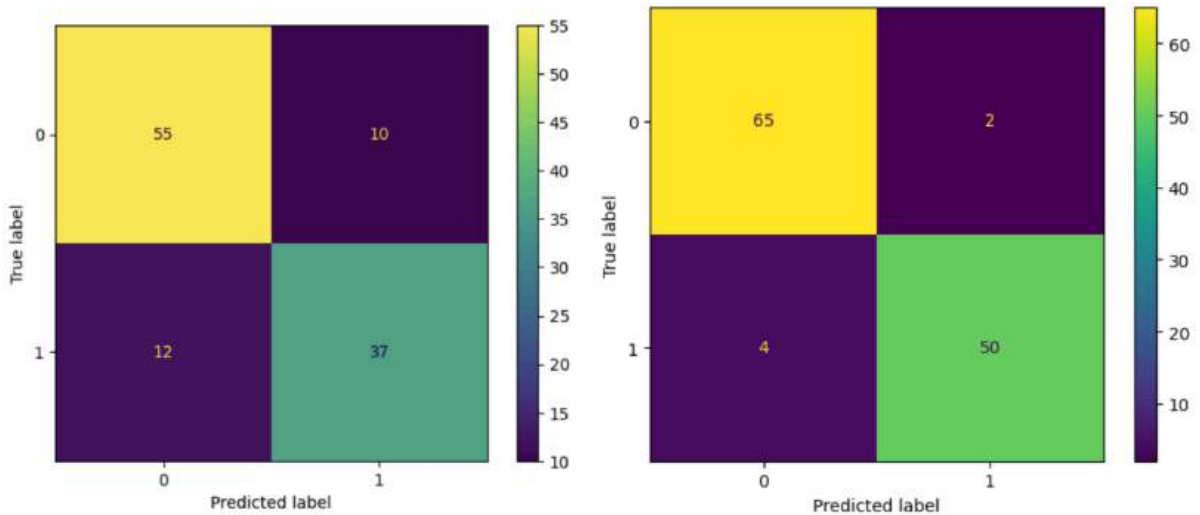
Table 2: Tesla Car Front EMI

AI Models	No Shielding Anomalies	Shielding Anomalies
Isolation Forest	6	7
Random Forest	22	6
Autoencoders	19	28



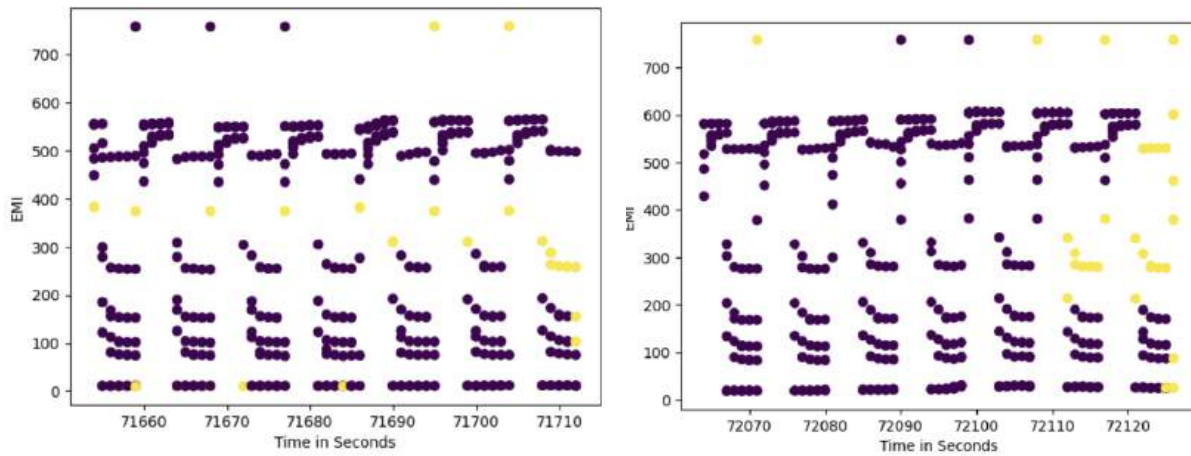
The results of EMI over time in seconds for the non-shielded data set (left) and the shielded data set (right).

Figure 14: Isolation forest



False positives are in the top right, false negatives are in the bottom left. True positives are in the top left, false positives are in the bottom right.

Figure 15: Random forest



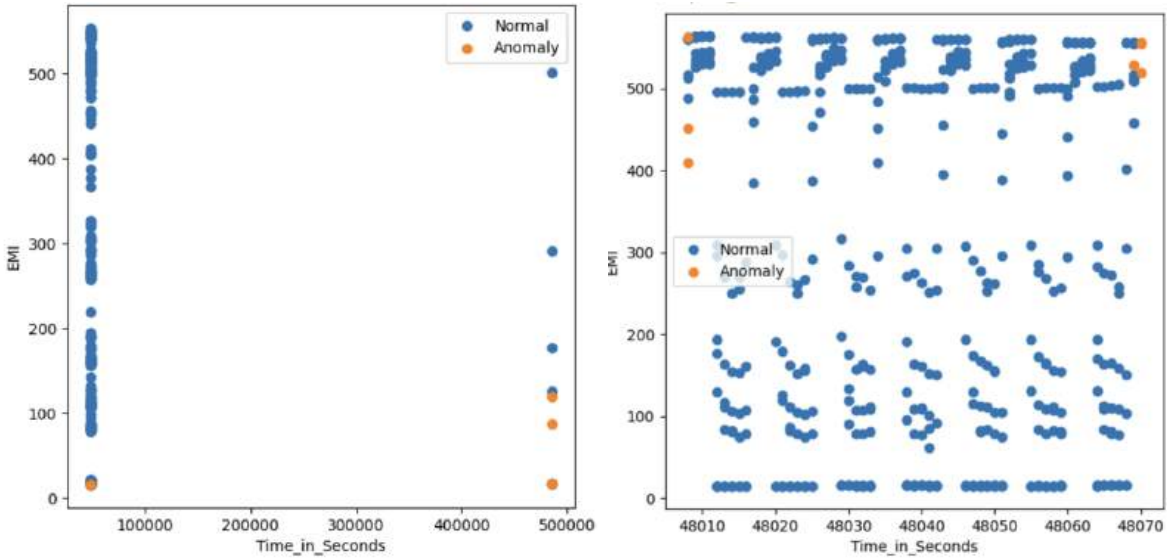
The results of EMI over time in seconds for the non-shielded data set (left) and the shielded data set (right).

Figure 16: Autoencoders Anomaly Detection Network

The models suggest that shielding introduces conditions that lead to irregularities in both blood pH and elevation readings. This is noted as more outliers are pointed out in general for the shielded tests, and blood pH, as well as elevation levels, tend to be skewed toward one side. This could imply faulty shielding setups.

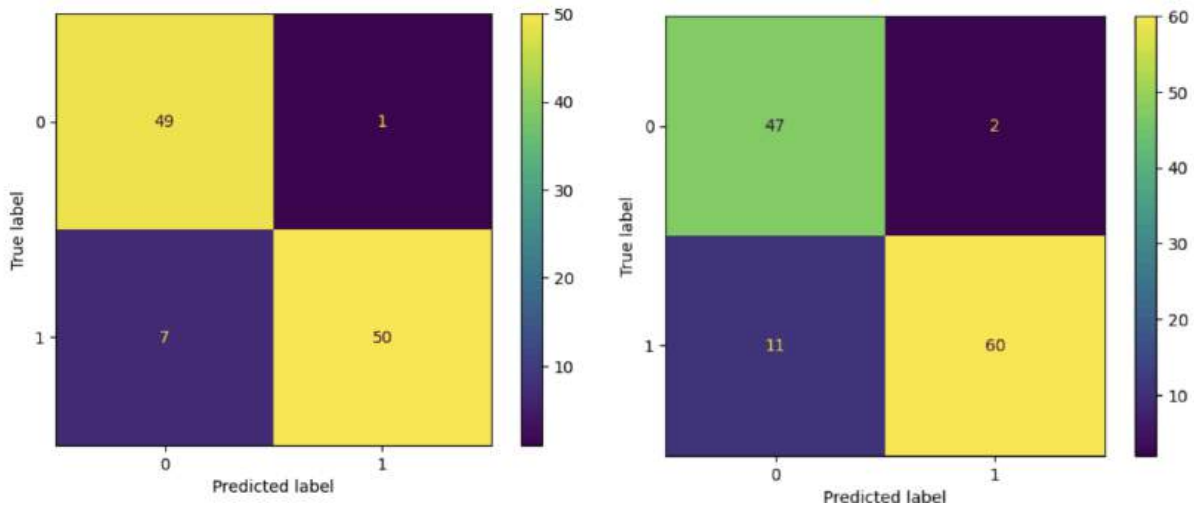
Table 3: WiFi Router EMI

AI Model	No Shielding Anomalies	Shielding Anomalies
Isolation Forest	6	7
Random Forest	8	13
Autoencoders	0	0



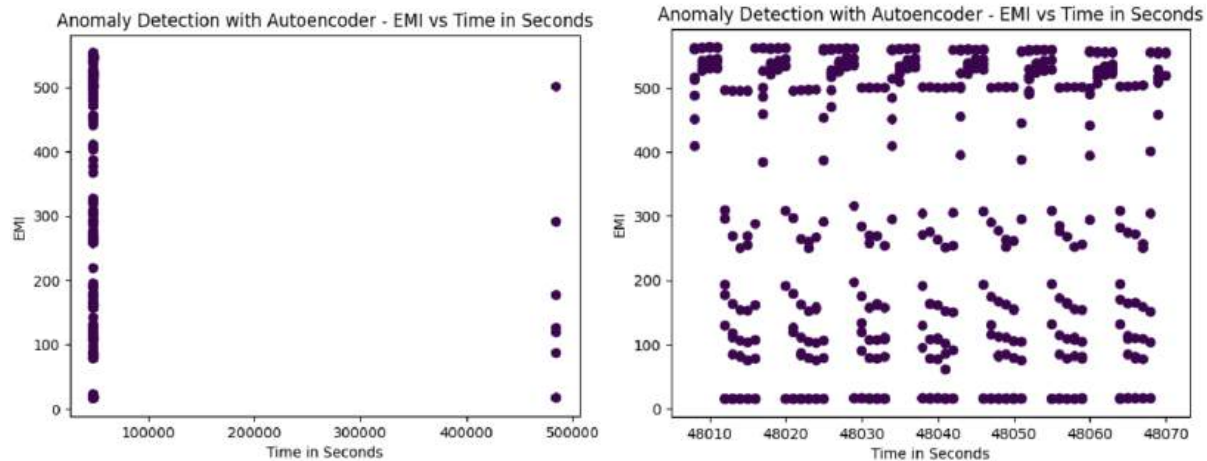
The results of EMI over time in seconds for the non-shielded data set (left) and the shielded data set (right).

Figure 17: Isolation forest



False positives are in the top right, false negatives are in the bottom left. True positives are in the top left, false positives are in the bottom right.

Figure 18: Random forest



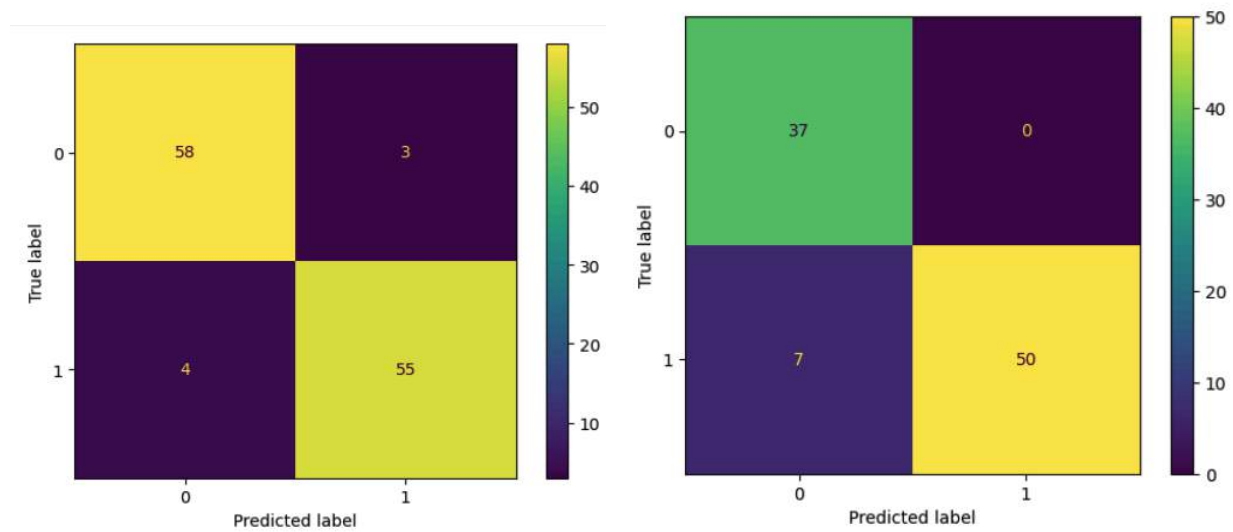
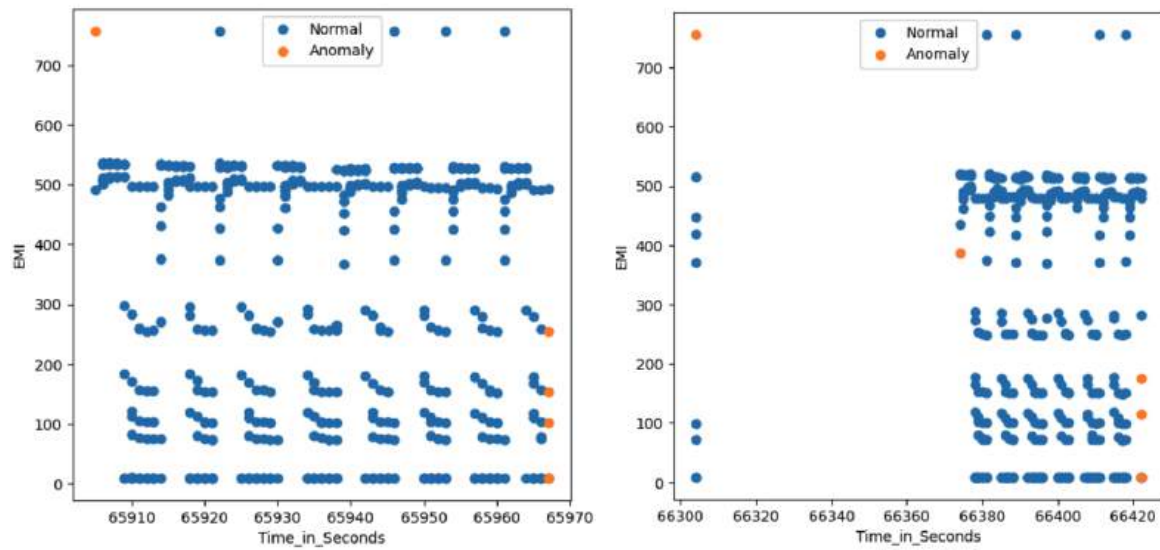
The results of EMI over time in seconds for the non-shielded data set (left) and the shielded data set (right).

Figure 19: Autoencoders Anomaly Detection Network

The graphs above show a very high skew to the left in the non-shielded data, and a relatively consistent spread in the shielded data. For blood pH and elevation above sea level, the models are skewed to the upper and lower ends, with outliers above an EMI of 300. For the WiFi data set, we see that shielding was a benefit as it improves consistency and reduces skew, but high EMI levels, anything greater than 300, still cause detectable disturbances. This is a good example of how shielding mitigates, but does not completely eliminate EMI. This data set has an unusual graph for both the autoencoder and isolation forest models for the non-shielded data sets. This could be explained by the threshold value being calculated to over- or under-fit the data.

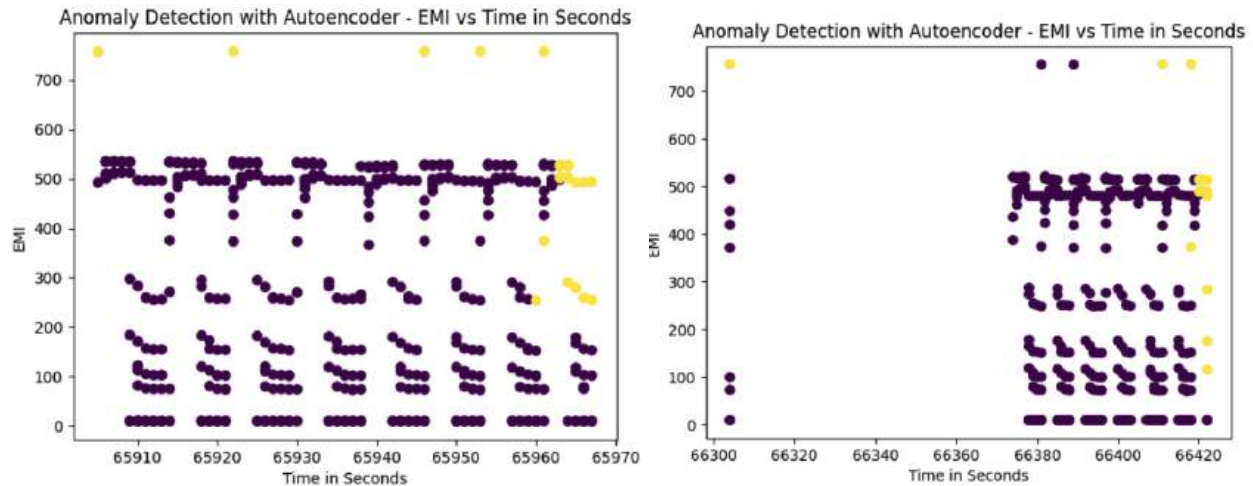
Table 4: iPhone Calling EMI

AI Model	No Shielding Anomalies	Shielding Anomalies
Isolation Forest	6	5
Random Forest	7	7
Autoencoders	18	13



False positives are in the top right, false negatives are in the bottom left. True positives are in the top left, false positives are in the bottom right.

Figure 21: Random forest



The results of EMI over time in seconds for the non-shielded data set (left) and the shielded data set (right).

Figure 22: Autoencoders Anomaly Detection Network

The findings for this data set indicate more outliers identified as time goes on, as well as a skew for the shielded data set towards the upper end of the graph. This could be due to the threshold factor once again. The trend in blood pH and elevation followed the ongoing trend of outliers skewed to the extremes. In essence, this indicates that as time goes on, more EMI is picked up, possibly due to cumulative EMI effects or threshold limitations. Furthermore, the upper skew of the shielded data implies that while shielding alters the data distribution, it does not eliminate the effects entirely— it could also shift or delay its effects.

Conclusion

Electromagnetic (EM) waves are ubiquitous, and they affect our daily lives without our knowledge. As technology further intertwines with modern society, the implications of EMI on critical medical devices such as pacemakers are all the more susceptible. This research examined the effects of EMI generated by common consumer devices on a simulated VVI pacemaker, including the integration of AI-assisted signal classification to improve detection and analysis. The results of the experiments further indicate that pacemakers may be distressed as a result of EMI from most devices, though not portraying the full scope of the interference, as the shielding of the pacemaker was not underneath human tissue, and not a complete emulation of pacemaker shielding. However, this study underscores the importance of robust shielding and fail-safe design in implantable medical technologies as technology is further integrated into present-day society.

Works Cited

- Worldwide, ForHearts. "Pacemakers and Electromagnetic Interference." Accessed August 10, 2025.
<https://forhearts.org/pacemakers-and-electromagnetic-interference-what-you-need-to-know>
- Kivrak, Elfidemiz, et al. "Effects of Electromagnetic Fields Exposure on the Antioxidant Defense System." Accessed August 12, 2025.
<https://doi.org/10.1016/j.jmau.2017.07.003>.
- Taranovich, Steve. "How Does EMI Affect the Human Body and Brain?" Accessed August 12, 2025.
<https://www.electronicdesign.com/technologies/power/whitepaper/21212534/electronic-design-how-does-emi-affect-the-human-body-and-brain>.
- Kaszuba-Zwoińska, Jolanta, et al. "Electromagnetic Field Induced Biological Effects in Humans." Accessed August 20, 2025. <https://pubmed.ncbi.nlm.nih.gov/27012122/>
- Engineering, Compliance. "How Harmful Is Electromagnetic Interference?" Accessed August 24, 2025.
<https://www.compeng.com.au/how-harmful-electromagnetic-interference/>.
- Erdogan, Okan. "Electromagnetic Interference on Pacemakers." Accessed August 31, 2025.
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1564060/>.
- Jeffrey, Kirk, and Victor Parsonnet. "Cardiac Pacing, 1960–1985: A Quarter Century of Medical and Industrial Innovation." Accessed September 3, 2025.
<https://doi.org/10.1161/01.CIR.97.19.1978>.
- Implantable Device Replacement Procedure. Johns Hopkins University. Accessed September 4, 2025.
<https://www.hopkinsmedicine.org/health/conditions-and-diseases/implantable-device-replacement-procedure>.
- "Cardiac Pacemaker." Yale Medicine. Accessed September 9, 2025.
<https://www.yalemedicine.org/conditions/cardiac-pacemaker>.
- "Pacemakers and Implantable Defibrillators." MedlinePlus. Accessed September 9, 2025.
<https://medlineplus.gov/pacemakersandimplantabledefibrillators.html>
- Goyal, Amandeep, et al. "Pacemaker Malfunction Due to Electric Blanket: A Rare Case of Electromagnetic Interference." Kansas Journal of Medicine. Accessed September 10, 2025. <https://doi.org/10.17161/kjm.vol15.18549>.
- Mulpuru, Siva K., et al. "Cardiac Pacemakers: Function, Troubleshooting, and Management: Part 1 of 2-Part Series." Journal of the American College of Cardiology. Accessed September 10, 2025. <https://doi.org/10.1016/j.jacc.2016.10.061>.
- "2 Easy, Accurate Ways to Measure Your Heart Rate." Mayo Clinic. Accessed September 10, 2025.
<https://www.mayoclinic.org/healthy-lifestyle/fitness/expert-answers/heart-rate/faq-2005979>.

Lak, Hassan Mehmood, and Amandeep Goyal. "Pacemaker Types and Selection." Accessed September 11, 2025.

<http://www.ncbi.nlm.nih.gov/books/NBK556011/>.

Beinart, Roy, and Saman Nazarian. "Effects of External Electrical and Magnetic Fields on Pacemakers and Defibrillators: From Engineering Principles to Clinical Practice." Accessed September 11, 2025.

<https://doi.org/10.1161/CIRCULATIONAHA.113.005697>.

"Pacemaker Insertion." Johns Hopkins University. Accessed September 12, 2025.

[https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/pacemaker-inserton](https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/pacemaker-insertion).

"What Is an EV Skateboard Chassis? Why It's Game-Changing." Fuel2Electric. Accessed September 13, 2025.

<https://www.fuel2electric.com/what-is-an-ev-skateboard-chassis-why-it-s-game-changing>

A Vision-Based Framework for Automated Biomechanical Analysis and Feedback in Cricket Fast Bowling By Subinay Kalya

ABSTRACT

In Cricket, objectively measuring fast bowling techniques is essential to improving performance and minimizing player injury risk. Reliant on motion capture systems or manual video review, traditional biomechanical analysis is out of the picture for most athletes. Using a dataset of annotated fast bowling videos, this study introduces an AI-based computer vision system that leverages pose estimation and ball-tracking models to extract key features from player movement. The system predicts necessary physical measures, such as hip-shoulder separation, stride length, and other angles, while accounting for the athlete's body and movements. The feedback part of the system relies on a model that draws from established biomechanical principles to turn the measures mentioned above into player-ready, understandable feedback that a bowler can use to improve performance and consistency. In testing, our framework produced accurate results across a variety of video settings and playing styles, matching prior research on efficient/inefficient biomechanical measures. Our framework offers a practical, affordable, and data-driven method to support Cricket fast bowling analysis and coaching.

I. Introduction

Quantitative fast-bowling biomechanics is a complex problem in both sport science and performance analytics. As with any movement of a fast bowler, there are many different components in the biomechanical chain (run-up, foot strike, hip-shoulder separation, knee-brace) that affect both the speed at which the ball travels and an athlete's susceptibility to injury. Historically, the biomechanics of each component have been evaluated using either high-speed motion capture systems or through manual annotation by a qualified biomechanist. Both methods provide excellent data; however, high-speed motion capture systems are very costly, must be used under laboratory conditions, and thus are typically not available to players and coaches outside of elite training environments. Therefore, all other bowlers rely on qualitative video feedback from their coaches, which cannot match the level of detail and accuracy required to develop evidence-based solutions to improve performance.

The need for automated, scalable tools to evaluate an athlete's biomechanics, specifically in cricket, is a broader issue in sports biomechanics. There are currently no tools that automatically analyze large volumes of real-match/practice bowling footage at low cost. This prevents athletes from learning how mechanical inefficiencies in their movements (reduced hip-shoulder separation, excessive lateral flexion, etc.) affect their ability to perform effectively during matches and/or increase their risk of injury. Additionally, when relying on expert manual evaluations, this variability among evaluators can limit the ability to replicate results across studies and thereby create a barrier to establishing standardized, data-driven measures for assessing player performance in cricket.

It's essential to develop solutions to address biomechanical challenges and improve athletic performance, and to provide equal access to biomechanical data at all levels of play (from grassroots to elite/professional). An easy-to-use, automated framework will provide athletes at all levels of the game with evidence-based biomechanical analysis through regular videos. In addition to improving performance, this type of analysis may help prevent injuries by identifying mechanical discrepancies that can lead to overuse injuries or improper energy and momentum transfer during the bowling delivery stride.

This paper introduces a computer-vision-based framework employing artificial intelligence (AI) to automatically extract and interpret the fundamental aspects of biomechanics for fast bowlers. This framework includes YOLO-Pose for joint localization, a customized ball-tracking model for event timing, and a feature-extraction pipeline for calculating kinematic measures scaled to an individual's body size. Kinematics used in this study include: Hip–Shoulder separation; Arm–Slot alignment; Release Height; Stride Length; Braking Impulse; and many more. Each of these metrics was selected for its biomechanical relevance and its relationship to performance, based on prior research. Finally, a biomechanical feedback model maps these calculated metrics to performance interpretations based on empirical research and known optimal mechanics.

Our concrete contributions are:

1. A computer vision pipeline combining pose estimation and ball tracking for automated biomechanical metric extraction; and
2. A structured framework for translating raw kinematic data into actionable coaching insights. This is achieved by mapping the bowler's computed metrics against optimal performance thresholds derived from biomechanics literature to generate specific, plain-language critiques and recommendations.

Collectively, these contributions form a foundation for scalable and accessible biomechanics analysis in cricket. While this study focuses on fast-bowling, the framework can be generalized to other sports domains involving high-speed, asymmetric motion, such as baseball pitching or javelin throwing.

II. Related Works/background

The studies related to the present research come from three critical areas: (i) the biomechanics of cricket and fast bowling, (ii) methods for analyzing motion in sport by vision systems, and (iii) systems for providing automated feedback and interpretation. Together they form the basis of this work, but they also show that no 2-D video–to–feedback end-to-end system is available.

2.1 Biomechanics of Fast Bowling

The mechanics involved in fast bowling performance have long been studied in

laboratory conditions. For example, Elliott et al. (2004) showed that ball release speed was influenced by the degree of hip–shoulder separation, bracing of the front leg, and trunk rotation (Elliott, Bruce, and Angus Burnett, 2004). In other studies, excess shoulder counter-rotation was found to be associated with the risk of lumbar spine injury (Senington, 2017), while hip and shoulder mobility were associated with technical efficiency (Felton et al., 2023). Although these studies have defined the normative mechanics involved, the need for multi-camera systems used with motion analysis software and force-plate systems restricts their availability to fast bowlers in the real world, suggesting that an alternative that can be field-deployed is required.

2.2 Computer Vision and Video-based Movement Analysis

Recent improvements in markerless motion capture have demonstrated the viability of 2D and Monocular 3D estimation as an accurate means of obtaining kinematic data in a “real-world” setting. Naik et al. (2022) examined computer vision applications in sport. They reported on pose estimation and trajectory tracking and highlighted the improved accuracy and reduced hardware costs for these applications (Naik et al, 2022).

Edriss et al. (2025) provided similar examples of low-cost vision sensors supporting biomechanical analysis of movement outside laboratory environments (Edriss et al, 2025).

Although there are many examples of computer vision applied to sports, there are relatively few specific to cricket. However, some currently available computer vision systems can provide both joint kinematics (how each part of the body moves) and ball flight metrics from a single camera recording.

2.3 Automated Coaching Feedback and Interpretation Systems

Sports-performance AI has increasingly focused on automating the generation of feedback. Li (2025) surveyed applications of generative AI in sport, noting its potential for real-time analysis and personalized instruction (Li, 2025). A small number of exploratory works have applied machine learning to bowling classification or performance clustering (Marwah, 2025), yet none link quantitative biomechanical metrics with structured natural-language feedback or injury-risk assessment.

2.4 Identified Gaps and Rationale

In summary, the current literature powerfully addresses (i) biomechanics of fast bowling in laboratory environments and (ii) technical feasibility of computer vision systems involved in measuring athlete movement. Less well addressed are (iii) systems that can be deployed in the field for extracting biomechanical parameters from ordinary video footage and (iv) automated creation of feedback protocols that relate the biomechanical parameters to sport-related skills and required coaching points. In this paper, we present an application that unifies laboratory biomechanics, computer vision analytics, and coaching intervention through the development of a pipeline which can be used to take single view video as input, produce output of joint position

and ball path information, derive a full list of bowling specific parameters from the previous inputs and create a final production of coaching derived feedback.

III. Methodology

3.1 Data Collection

To understand how artificial intelligence can evaluate and analyze fast-bowling mechanics, we utilized a set of 12 videos from both amateur and professional bowlers. The dataset consisted of seven videos from a back view of the bowler and five from a side-on view near the stumps. All of the captured clips used single-camera setups.

The selected videos ranged in level of play and delivery types. Amateur videos were captured from club-level bowlers, including me(the author), a U19 fast bowler for the Seattle team. We also used professional footage sourced from publicly available clips of Premier First Grade Cricket in Australia, featuring international bowler James Pattinson. We equipped this combination to introduce natural variability in skill, speed, and bowling technique to the model, all of which are key factors in creating a robust model.

Each video angle offered important insights into the bowling motion. Side view recordings highlighted stride length, trunk flexion, and arm-leg coordination. While back-view recordings emphasized shoulder-hip separation, body alignment, and ball trajectory. All features are closely associated with performance and injury risk.

Videos were included if they met the following criteria:

- i) the complete delivery sequence was visible from run-up through ball release;
- ii) both bowler and stumps were unobstructed throughout the motion;
- iii) the ball remained visible from release to impact at the wicketkeeper or stumps; and
- iv) video quality met a minimum resolution of 720p or higher.

The collected videos differ in terms of resolution, lighting, and frame rate. While some came from competitive match settings, others were recorded in controlled practice settings. The aim is for the model to adapt to the messy, real-world conditions in which bowlers are usually recorded, thanks to this variability, which also improves ecological validity despite posing challenges such as motion blur and uneven backgrounds.

3.2 Data Preprocessing and Model Training

All videos were stored in MP4 format at a minimum resolution of 790p (preferably 1080×1920), with clips cropped to isolate the bowler's run-up, delivery stride, and follow-through. Raw footage from videos was trimmed into clips that started in the middle of the bowler's run-up and ended with the ball either crossing the stumps or reaching the keeper. Each segment was resized to 640×640 pixels, aligning with the input requirements of the YOLO(explained below) framework. Although there was a difference in frame rates between clips, this was intentionally kept to reflect the differences in real-life recording situations.

3.2.1 Object Detection Framework

To extract biomechanical data from the dataset, we used a pose/object detection model. We decided to use YOLO (You Only Look Once), a learning-based object detection environment with various models available. We used Yolo for its real-time object localisation and classification in video frames. In contrast to standard two-stage object detectors, YOLO essentially comprises one neural network predicting bounding boxes and class probabilities at the same time. This produces improved speeds and accuracy, which give it efficiency and robustness, making it applicable to sports biomechanics, where fast movement and small objects such as cricket balls must be tracked successfully.

We used YOLO and its YOLO-Pose model as the foundation of our pipeline due to their proven performance in human pose estimation and their adaptability through fine-tuning. This makes it a viable model for deployment in all cricketing environments without the need for state-of-the-art equipment.

3.2.2 Ball Detection Model

Since a cricket ball is not one of YOLO's default object classes, a custom-trained version of YOLOv8 was developed to find cricket balls. There were approximately 486 manual bounding box annotations on frames from videos of amateur and professional cricket that contained a ball. The data were split into training (70%), validation (30%), and test (10%).

No synthetic augmentation was used in the initial phase; instead, we retained the natural variability in lighting, motion blur, and trajectory. The trained model outputs bounding box coordinates per frame, from which the ball center point was extracted as the tracking reference.

3.2.3 Pose Estimation Model

After being pre-trained on general human pose data, a different YOLO-Pose model was optimized using approximately 705 frames from the project's database. YOLO-Pose has been shown to perform reliably on human movement estimation (Ultralytics, 2024; Naik et al., 2022; Baily, 2020). Some of the features observed in the collection of bowling frames included exaggerated forward trunk flexion, occluded limb positions, and prolonged arm rotation. The bowler's arms, hips, ankles, and shoulders were the primary focus of model fine-tuning, as these keypoints are biomechanically significant in fast bowling mechanics.

KeyPoints:

Nose, Left Eye, Right Eye, Left Ear, Right Ear, Left Shoulder, Right Shoulder, Left Elbow, Right Elbow, Left Wrist, Right Wrist, Left Hip, Right Hip, Left Knee, Right Knee, Left Ankle, Right Ankle.

3.2.4 Training Procedure

Utilising the default YOLO hyper-parameter settings, both models were trained using stochastic gradient descent with a learning rate of 0.01 and a total of 60 epochs. Performance was

communicated via loss graphs and precision metrics over the validation dataset, through the mean average precision (mAP) at intersection over union (IoU) thresholds ranging from 0.5 to 0.95. Model training was monitored for overfitting, ensuring that the best weights were selected for downstream inference.

After being exported in PyTorch format, the completed models were incorporated into the primary analysis pipeline, which automatically extracts pose and ball-tracking information from video data.

3.3 Pose & Ball Estimation Pipeline

To translate raw cricket videos into biomechanical data, we created a pipeline of Python scripts that combined the outputs of the two aforementioned detection models: YOLO_Pose for pose estimation and our custom-trained YOLO_v8 ball-detection model.

YOLO_pose generated 2D x_y skeletal keypoints for the bowler, while the YOLO_v8 ball model localized the cricket ball through similar x_y keypoints (COCO format). We use these outputs to map body motion and ball trajectory in space using general calculations, enabling subsequent calculations of important fast-bowling metrics.

-
- combined_pipeline.py — handles video ingestion, pose and ball inference, calibration, tracking, overlay generation, and output creation (CSV, JSON, MP4).
- event_detect.py — identifies key bowling events such as back-foot contact (BFC) and front-foot contact (FFC) by analyzing smoothed ankle trajectories relative to the release frame.

3.3.1 Inference and Data Capture

Using OpenCV, each video was evaluated at the video's native frame rate ~ 25-30 fps. The models produced keypoint coordinates as mentioned above. In addition, the pose model produces a skeletal outline of the player, and the ball model produces a single bounding box and a red dot identifying the cricket ball for each frame. The identity of the bowler across frames was retained using a simple nearest-neighbour association, tracking the detected person closest to the previous location.

Automatically, the outputs of the models were sorted into 4 files:

- (i) an annotated MP4 visualizing the detected skeleton, ball path, and field cues;
- (ii) a CSV file containing per-frame ball positions;
- (iii) a CSV file containing per-frame keypoint positions; and
- (iv) a JSON summary storing detected events, timing, and velocity metrics. This structured output enabled both quantitative and qualitative analysis of each delivery.

3.3.2 Temporal smoothing and occlusion handling-

We chose to use a hybrid approach around the release point to improve cricket ball tracking. The ball becomes smaller, faster, and is occasionally occluded in the video frames. For a short period of time, the system uses the wrist position of the bowler at release as a proxy for the ball. Therefore, we check if the wrist keypoint is visible before release. The detections are fused with wrist motion predictions using a constant-velocity Kalman filter, which smooths jitter and blanking after the ball detector and provides confidence (Kalman, 1960). A simple gating mechanism is then used, which favors detections close to the predicted positions and limits false positions resulting from background disturbances.

The CSV file contains all points in the video where the ball was detected. These points are labeled as either model predictions, sensor measurements, or wrist proxy. These will be used to calculate the velocity and ball trajectory.

3.3.3 Pitch cues and metric scaling.

To convert data to real units, the code uses an in-frame calibration technique: the release frame is analyzed, and the bowler's height is measured in pixels. This value is divided by the player's input height to yield a conversion factor in meters per pixel (Edriss et al., 2025). This conversion factor is stored in the JSON summary and used throughout the following processes when needed.

To find the stumps (vertical clusters) and the popping crease (horizontal line) in each video, a lightweight detection process was developed. These contextual references are critical for calculations.



Figure 1: Image showing Stumps and Popping Crease on the cricket pitch

3.3.4 Automated Event Detection

To calculate a player's features and metrics, it is important to know when a set of key events occurs. To help with this, a Python script, `event_detect.py`, was created.

Here are the key events (Portus et al., 2004; Worthington et al., 2013):

- **Back Foot Contact(BFC)**- BFC is the last time the player's dominant side ankle touches the ground before ball release

- **Front Foot Contact(FFC)** - FFC is the step the bowler takes with their non-dominant leg milliseconds before ball release.
- **Ball Release** - The frame when the bowler releases the ball. This is detected automatically when the distance between the ball and the wrist abruptly increases.



Figure 2: 3 Images showing BFC, FFC, Ball-Release

Once the release is known, `event_detect.py` searches backward through the keypoints' frames to find ground-contact peaks in ankle trajectories. Since keypoints are detected and shown as x,y coordinates, the last high y-coordinate point on the back-ankle marker is considered the "Back Foot Contact", and the previous high y-coordinate point on the opposite ankle is marked as "Front Foot Contact. Both these events are searched for within realistic windows before release. These events occur during the bowler's release stage of the delivery and are essential for deriving additional metrics and conducting mechanical analyses.

The summary JSON stores these event times (in seconds and frame numbers) along with which foot served as the back foot for that delivery.

3.3.5 Speed and Flight Computation

Utilizing the pixel-to-meter scale, ball speed at the point of release is calculated from the ball path immediately after release. By computing the amount of time the ball takes to pass the wickets, we compute the average flight speed. Both metrics are reported in m/s and km/h. These values, combined with the timing of BFC and FFC, provide insight into the bowler's rhythm and efficiency during the delivery stride (Elliott et al., 2004).

3.3.6 Outputs and Generalization

The pipeline works with side-view footage as well, as long as a fixed reference, like the stumps or a calibrated object, is visible, even though the majority of the experiments were

conducted with back-view recordings. Raw video and biomechanical interpretation are transparently connected through outputs, including annotated MP4s, CSV logs, and a structured JSON summary. The subsequent step in the system uses these outputs directly to extract motion data and generate mechanical analyses and quantitative metrics.

3.4 Feature Extraction & Metrics

Following the pose and ball tracking models, the data is essentially the x,y coordinates of the ball and joints over the period of the video. These coordinates need to be transformed into meaningful measurements and metrics that explain a bowler's bowling action. These measurements are derived from calculations performed by a metric-extraction module written in Python scripts, which convert thousands of numbers into a clear picture of how the athlete operates. The derivations for each calculation are shown in [Appendix A](#).

The following metrics were extracted for each delivery:

- **Lower-limb mechanics:** back-knee angle at BFC, front-knee angle at FFC, stride length, and BFC contact time.
- **Trunk and upper-body motion:** hip-shoulder separation, lateral trunk flexion, and release height.
- **Temporal and velocity features:** run-up speed, wrist speed, and estimated ball release speed.
- **Positional descriptors:** landing type (toe-first/flat), arm position at FFC, and event timing (BFC → FFC → release).

These features collectively quantify the bowler's efficiency, pace, sequencing, and injury risk profile, serving as input to the feedback model (Elliott et al., 2004; Felton et al., 2023).

3.4.1 Preparing the data

However, before any measurements are taken, the data is cleaned; frames that contain uncertain landmarks are ignored, and short averages are applied to smooth out jitter from camera movement.

3.4.2 How it all connects

The system associates each calculation to the three key events in the bowling action: Back-Foot Contact, Front-Foot Contact, and Release. These moments serve as reference points for measuring angles from detected joints, computing distances, and deriving speeds.

3.4.3 Interpretation Framework

Collectively, these metrics reconstruct a bowler's kinetic sequence:

1. **Approach phase:** Run-up speed and impulse stride quantify approach energy and preparatory momentum.

2. **Delivery phase:** Back-knee angle, jump height, and hip–shoulder separation describe power transfer and rotational mechanics.
3. **Release phase:** Lateral flexion, release height, and release/flight speeds summarize delivery efficiency and control
4. **Follow Through Phase:** Front knee brace, trunk bend, knee drive, interpret the preservation of momentum and energy through the action.

(Glazier & Wheat, 2013)

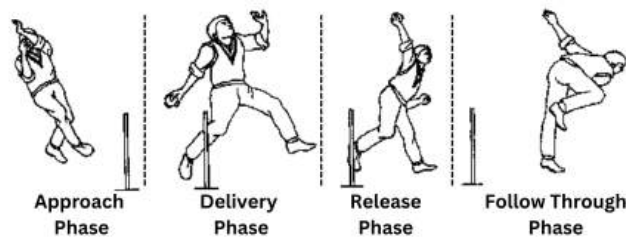


Figure 3: Illustrations of the phases of a fast bowling action

Capturing and analyzing these phases with a single camera, a minimalistic setup, and computer vision enables our system to provide biomechanical feedback across all levels of cricket. The resulting outputs can guide coaching correction and associate mechanic patterns with efficiency, injuries, and performance.

3.5 Interpretation and Feedback Model (LLM-Based Analysis and Feedback)

3.5.1 Purpose

Although the numerical metrics derived show joint angles, timing, and speeds, they cannot explain why or how a bowler moves/bowls inefficiently. Translating biomechanical outputs into coaching is imperative, which is why our feedback model was created to interpret the metrics in the context of cricket-specific mechanical knowledge. By identifying bowling styles (side-on, front-on, mixed) and bowling dominance (hip- or knee-driven), it provides targeted, personalized feedback.

3.5.2 Rationale

Metrics such as hip–shoulder separation or release height lack meaning without context. For example, a 20° separation may be adequate for a front-on bowler but sub-optimal for a side-on bowler.

The feedback model will provide this missing layer of explanation by recognizing patterns across metrics and associating them with previously conducted research to offer clear feedback.

3.5.3 Description of the Model & Process

Our Feedback Model is a hybrid system combining symbolic (rule-based) inference, light classification, and a language model for narrated feedback.

The system is:

1. **Feature normalization & banding:** metrics are scaled relatively to layer height and categorized as Subpar, Optimal, Above, Dangerous, or Uncertain using ranges that derive from expert literature
2. **Derived features & scoring:** utilizing available metrics, the system derives a set of feature vectors per delivery. (E.g., Peak separation, Event of peak, shoulder deviation, etc.)
3. **Alignment & dominance inference:** we developed heuristic scorers that assign likely labels for bowling styles(side-on / front-on / mixed) and bowling dominance(hip/knee).
4. **Rule-based logic + knowledge retrieval:** the model utilizes an internal knowledge base formatted in YAML files that contain: encoded normalized values, fault & cause mapping, drill libraries, and risk associations.
5. **Narrative Feedback Generation via LLM:** findings, alignment/dominances, and retrieved knowledge are formatted as structured input to a large language model. The JSON scheme contains the following: i)style and dominance; ii) overall_assessment; iii) a list of findings with justification, risk, and actionable tips; iv) a short list of drills; v) watchouts (uncertainties or missing data); vi) confidence and data_gaps

To ensure that all responses are consistent and factual, a validation step via JSON Schema is utilized. In the case that validation fails, the system will re-prompt the LLM for a new output. The LLM is input structured evidence and features rather than plain text. This allows its feedback to be grounded in its domain knowledge rather than open-ended generation.

3.5.4 Implementation Details

The language component of the model is a schema-constrained LLM. We opted to use Llama-3 8b-Instruct, an open-source component that is served via vLLM for low-latency inference. At run-time, the model receives the following: (i) the derived feature vector and metric bands (JSON), (ii) rule-engine hits (Python rules over banded metrics), (iii) Short snippets of biomechanic knowledge are retrieved from the structured YAML knowledge base and indexed with BGE-large embeddings in a FAISS vector store. Orchestration is handled by LlamaIndex, which utilizes RAG graphing and reranking, and the prompt applies a strict JSON schema. The following keys are used: style, dominance, overall_assessment, findings[], drills[], watchouts[], and confidence.

Outputs are validated with Pydantic v2 (+ jsonschema); on validation failure, we reprompt with the validator error only to correct the structure (never to invent new numbers). The service is exposed as a FastAPI endpoint (POST /feedback), with determinism tuned by

temperature ≈ 0.2 and top_p ≈ 0.9 .

3.5.5 Why RAG, Not Training?

Retrieval Augmented Generation provides a better alternative to training (fine-tune) in this case, due to its ability to keep the model focused on the specific biomechanical aspects of Cricket and the defined normative ranges. The model can generate verifiable recommendations by citing snippets it retrieves and rules tied to measurable features. This makes the model more reproducible, less prone to hallucinations, and computationally inexpensive enough to be used for each delivery.

3.5.6 Knowledge Base and References

The Feedback Model draws on established fast-bowling research to define normative ranges and mechanical relationships. The research used is all the sources cited in this paper.

3.5.7 Computational Logic

Each delivery's metrics are transformed into interpretable features and probabilistic scores:

3.5.7.1 Additional Derived Features

To accurately derive aspects of a bowling action like style, dominance, and stress factors, the following additional features were derived from the pose-estimation outputs and event timings (BFC, FFC, and Release): hip-shoulder separation(at BFC, FFC, and Release), arm slots, and braking impulse time at release.

From these, several secondary variables were computed:

- **Hip-Shoulder Separation-Peak:** identifies when trunk rotation was maximized (BFC, FFC, or Release)
- **Shoulder Line vs Pitch Release:** measures whether the torso remains closed (side-on) or open (front-on) at release.
- **Arm Slot:** represents the angle between the shoulder, elbow, and wrist compared to a general vertical at the point of release.
- **Braking Impulse Time:** identifies timing where negative values indicate early braking, positive values indicate late braking relative to release.

3.5.7.2 Style and Dominance Scoring

To derive bowling style from the extracted metrics and features, we used a heuristic scoring method that converts these features into normalized percentages for style and dominance (Portus et al., 2004; Elliott et al., 2004).

- **Side-on:** If the hip-shoulder separation angle shows significant early separation $\geq 25^\circ$ at BFC, shoulder closure $\geq 15^\circ$ at BFC, and a vertical arm slot $\leq 25^\circ$ at release

- **Front-on:** If the hip-shoulder separation angle shows slight separation $\leq 15^\circ$ at BFC, square shoulders $\leq 10^\circ$ at BFC, and a lateral arm slot ($\geq 30^\circ$) at release.
- **Mixed:** If hip-shoulder separation at BFC is almost negligible or counter-rotation occurs

To derive hip or knee dominance in a bowler, the system utilized temporal and angular relationships between features and keypoints:

- **Hip-dominant:** if the bowler shows early braking, strong early separation, knee-collapse at BFC > 20 degrees, BFC contact time < 0.18 ms, and moderate lateral flexion.
- **Knee-dominant:** if the bowler shows late braking, slight separation, knee-collapse at BFC < 20 degrees, BFC contact time > 0.15 ms, and higher lateral flexion.

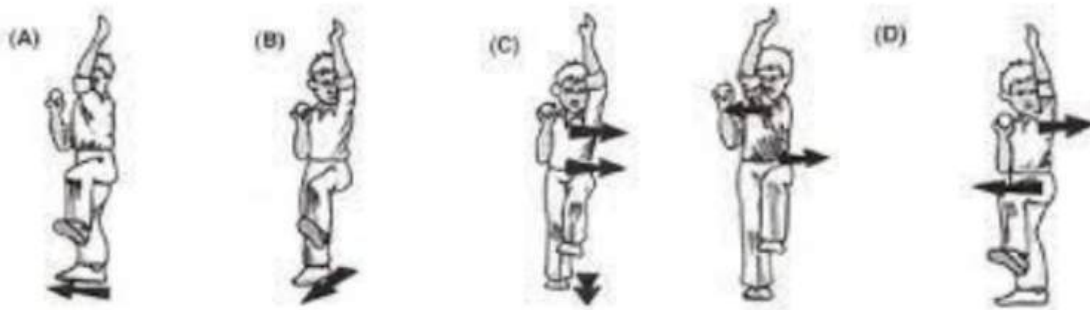


Figure 4: illustrations of (A) side on; (B) mid-way; (C) front on, and (D) two mixed actions (taken from Portus et al., 2004).

3.5.7.3 Rule Application

Each metric's categorical band (Below / Optimal / Above/ Dangerous/ Unown) and the bowler's style/dominance are evaluated against a compact rule base, an example of this was stated previously in 3.5.7.2.

Each rule produces a structured finding with the metric, status, reasoning, risk flag, corrective drill/action, and priority.

3.5.7.4 Structured Output

All results are compiled into a validated JSON structure by a final layer. Each JSON contains: Overall_assessment, Style, Dominance, Findings[], Drills, Watchouts, Confidence, Full List of Metrics

3.5.8 Output and Role

The feedback model converts large amounts of quantitative data into a more easily understandable biomechanical (movement-based) profile of the bowler's movements. This model will allow both athletes and coaches to see movement mechanics in the same way an expert bowling coach or analyst would: identify what is working and what is not, and understand how to improve inefficiencies.

IV. Results

In total, eleven videos of right-arm fast bowlers, comprising seven indoor practice sessions and four outdoor match recordings across three different bowlers and five different camera setups, were analyzed. The final, trained pipeline consists of a YOLOv8s-pose model for player keypoint detection and a YOLOv8n model trained on custom ball data for ball tracking (Ultralytics, 2024). Both models were evaluated on their respective validation datasets, which contain real-world footage. Each experiment was conducted on an NVIDIA RTX 3050 GPU (4GB VRAM || 8GB RAM) with an input resolution of 960px.

4.1 Detection and Pose Estimation Accuracy

Our ball detection model, YOLOv8n-ball, was trained twice to evaluate confluence stability and parameter sensitivity. Trained for 100 epochs, the first run achieved a mean average precision($mAP@0.5$) of 0.89 and $mAP@0.5:0.95$ of 0.49, along with a precision of 0.96 and a recall of 0.82. The second run used a slightly smaller dataset, trained for 60 epochs, and modified the augmentation parameters somewhat; it achieved $mAP@0.5$ of 0.75 and $mAP@0.5:0.95$ of 0.34, along with precision of 1.00 and recall of 0.61.

Throughout training and validation, both models showed a consistent downward trend in loss curves across all components(box, classification, and DFL). This indicates stable optimization and the apparent absence of overfitting. The figure below, Figure 7, shows the learning curves for both training and validation runs, illustrating progressive improvements in accuracy and precision.

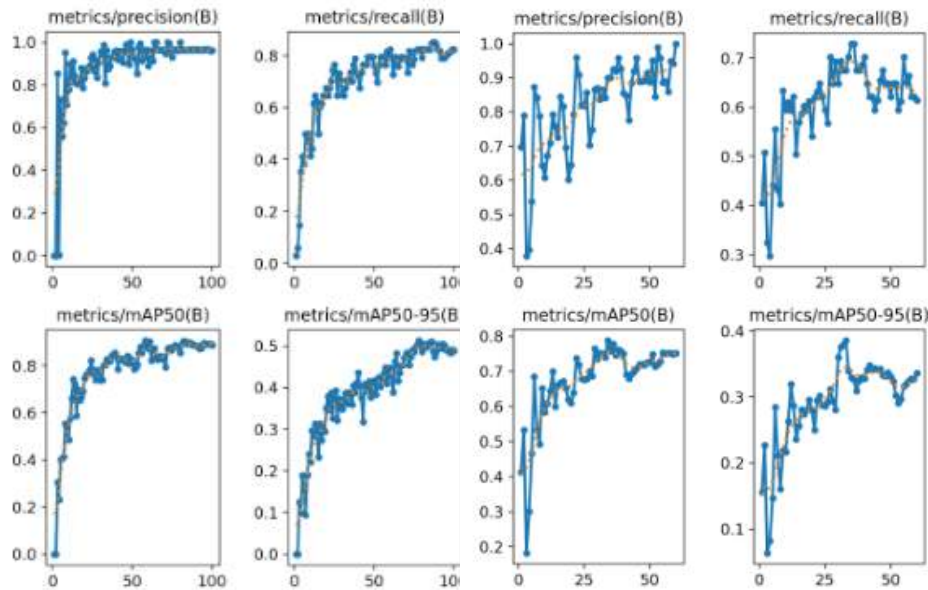


Figure 5: The Left four Graphs are of the first training. The right four graphs are of the second training run

Across all evaluated video sequences, the pose model (YOLOv8s-pose) illustrated reliable key-point localization, accurately predicting limb and trunk segments even under occlusion and motion blur, even without fine-tuning on cricket-specific videos. Misplacements at key points were extremely minor and occurred only during extremely fast, explosive movements, yet joint estimates at these events remained precise.

Ball tracking detections had occasional frame-level interruptions before release and immediately after the bounce, but post-release detections were relatively stable. To mitigate pre-release flickering, a wrist-anchored strategy was implemented where the wrist was used as a proxy for the ball before release. This ensured continuity between the final pre-release frame and the initial post-release trajectory.

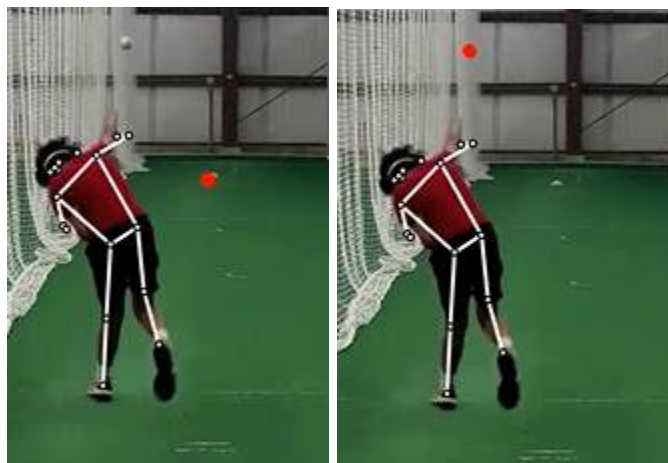


Figure 6: Before and after images indicating ball tracking improvement.

Overall, both the pose and ball models achieved strong accuracy and temporal consistency across diverse lighting and camera conditions, providing a stable foundation for the downstream event-detection and biomechanical-metric stages.

4.2 Event Detection Reliability

To separate a delivery into meaningful phases, it is essential to accurately predict key events such as Back Foot Contact (BFC), Front Foot Contact (FFC), and Ball release. To test the reliability of event detection, the pipeline was evaluated on four distinct videos, differing in setting, bowling style, and video setup. We compared frame indices to manually labeled truths (Table 1). Each video was recorded at 30 fps.

VIDEO NAME	TYPE	BFC (ACTUAL)	BFC (PRED.)	FFC (ACTUAL)	FFC (PRED.)	RELEASE (ACTUAL)	RELEASE (PRED.)
Subinay	Indoor	30	30	35	35	38	38
James Pattinson	Outdoor	66	66	71	72	74	74
Jinay	Indoor	37	38	43	43	46	46
James Pattinson 2	Outdoor	52	52	58	58	60	60

Table 1: Comparison of frame indices to truths of events.

When comparing predictions to the actual timestamps for the four test videos, the system obtained a mean absolute deviation across all 3 events for the four videos of only 0.17 frames, or 6 milliseconds, confirming high precision.

4.3 Computed Biomechanical Metrics

The third step in the pipeline, the biomechanical extraction module, produced kinematic measurements for each selected test video. These measurements are a variety of joining angles, temporal events, distances, and speed estimates.

VIDEO	J. PATTINSON 1	J. PATTINSON 2	J. DOSHI	S. KALYA 1	S. KALYA 2	S. KALYA 3
Camera View	Back Angle Outdoor	Back Angle Outdoor	Back Angle Indoor	Side-On Outdoor	Side-On Outdoor	Back Angle Indoor
Back-Knee Angle at BFC (°)	160.3	157.4	140.0	163.4	166.9	152.6
Front-Knee Angle at FFC (°)	175.6	173.5	173.6	169.9	174.7	174.6
Hip-Shoulder Separation (°)	39.8	44.5	24.9	0*	0*	47.6
Lateral Flexion (°)	34.9	35.5	23.2	32.8	33.7	29.7

VIDEO	J. PATTINSON 1	J. PATTINSON 2	J. DOSHI	S. KALYA 1	S. KALYA 2	S. KALYA 3
Stride Length (m)	0.09	0.04	0.14	0.97	1.15	0.16
Release Height (m)	1.83	2.14	2.19	1.63	1.81	1.80
Run-up Speed (kph)	2.6	2.3	2.5	15.9	19.7	4.0
BFC Contact Time (ms)	200	200	167	100	167	176
BFC Landing Type	Toe-first	Toe-first	Flat	Toe-first	Toe-first	Toe-first
Arm Pos @ FFC	Below shoulder	Above shoulder	Below shoulder	Below shoulder	At shoulder	Below shoulder
Release Speed (mph)	13.3	15.8	8.7	72.0	73.8	13.0

Table 2: Core metrics that were predicted for five videos

4.3.1 Angle-Dependent Metric Derivability

Camera perspective strongly influences metric reliability. Back-angle views (e.g., Pattinson 1–2, J. Doshi, Kalya 3) proved superior for analyzing horizontal and frontal-plane motion, enabling accurate estimation of hip–shoulder separation, lateral flexion, knee alignment, and full rotational sequencing—critical factors in assessing pace generation and dominance type (hip-, knee-, or mixed-dominant) (Elliott et al., 2004; Felton et al., 2023). However, because the ball’s trajectory aligns with the image’s y-axis, these views tend to underestimate ball velocity.

Conversely, side-angle views (e.g., Kalya 1–2) captured translational motion along the x-axis, allowing precise computation of stride length, release height, run-up speed, wrist speed, and ball velocity. Yet, the sagittal projection caused keypoint overlap, reducing rotational metrics such as hip–shoulder separation toward zero.

These results highlight the dependence on geometry that exists within biomechanical analyses. Each camera setup provides a partial yet complementary description of the delivery movement, indicating that multi-camera methods might be necessary to develop a fully three-dimensional reconstruction. While the current method was developed to adapt to this limitation, it still produced biomechanically valid measures at each camera angle.

4.3.2 Quantitative Patterns and Trends

Across five analyzed deliveries, the computed metrics showed consistent biomechanical patterns in these metrics: Hip-Shoulder Separation, Lateral Flexion, Front-Knee Flexion, Back-Knee Flexion, and Velocity metrics.

Patterns that align with the mechanics of hip-dominant actions were evident, as hip-shoulder separation averaged $31 \pm 17^\circ$ in side-on bowlers (e.g., Pattinson, Kalya), where delayed shoulder rotation and greater pelvis torque increase energy transfer. Contrastingly, lower hip-shoulder separation angles were seen in knee-dominant bowlers (e.g., Doshi), aligning with more front-on actions.

General practice shows side-on bowlers with lateral flexion $33\text{--}50^\circ$ while front-on bowlers exhibit lateral flexion angles ranging from $18\text{--}30^\circ$ (Worthington et al., 2013; Portus et al., 2004). Lateral flexion readings from the five selected videos showed an average of 34° for side-on bowlers and 23° for front-on bowlers. Front-knee extension at FFC remained strong across all bowlers, while back-knee angles had higher collapses for front-on bowlers ($\sim 140^\circ$) compared to side-on bowlers ($\sim 160^\circ$).

Collectively, these trends reaffirm that side-on, hip-dominant actions emphasize trunk rotation and torque generation. In contrast, front-on, knee-dominant styles rely more on linear drive and front-leg bracing.

4.3.3 Interpretation and Reliability

Manual inspection of two reference deliveries confirmed that automated estimates were within $\pm 8^\circ$ for angular variables and ± 0.1 m for spatial metrics.

Side-angle derived release speeds (72–74 mph) aligned closely with radar-gun readings, while back-angle speeds were ignored due to projection distortion.

Across all perspectives, wrist-speed and knee-angle values displayed strong consistency, demonstrating the model's ability to extract both proximal and distal kinematics from 2-D video.



Figure 6: James Pattinson's estimated Lateral Flexion was 35.5. The manual Calculation shows 43.8. A difference of only 8.3 degrees.



Figure 7: Subinay Kalya's estimated Knee Angle @FFC was 174.7. The manual Calculation shows 172.8, a difference of only 1.9 degrees. His Stride Length was estimated at 1.15m; converting 244.47px to meters yielded 1.23m.

4.4 Feedback Model Performance

Feedback is being delivered through a structured, schema-restricted language module where the biomechanical assessment and coaching advice will be derived from the biomechanical features computed in the analysis phase. The output of the model will be provided as a JSON structure that contains categorical fields for Style, Dominance, Findings, Drills, Watchouts, and Confidence. Representative outputs for two deliveries are shown below, illustrating both the interpretive depth and the model's alignment with biomechanical logic.

4.4.1 Feedback Examples

James Pattinson (Semi-Side-On, Back Angle)

The model identified a hip-dominant action with a strong vertical component, correctly associating a 157° back-knee flexion at BFC with controlled load absorption and a 173° front-knee extension at FFC with effective bracing for energy transfer. A -40° lateral flexion was noted as excessive contralateral tilt, beneficial for release height but potentially stressful on the lumbar region. Drills recommended by the model to enhance rotational timing and tendon stiffness included a single-leg jump squat and a medicine ball throw. The high confidence (.88) indicates a strong relationship between the quantitative and qualitative results of this study.

Subinay Kalya (Side-On, Mixed Action)

During this trial, the model identified the bowler as a knee-dominant mixed bowler with an efficient sequence. It noted a front-knee extension of 169.9° and hip-shoulder separation of 48.8° , describing the latter as “functional but below elite benchmark.” Lateral flexion of 32.8° was correctly classified as aggressive yet biomechanically valid for a side-on action. The system's specific drill/exercise suggestions of "resisted run up sprints", "paloof press core work" directly relate to increasing the efficiency of the kinetic chain and controlling lateral flexion in the bowler's body. Even at a low run-up speed (15.9 kph), the model determined that the bowlers had an efficiency in transferring their energy, as indicated by a wrist speed of 19.6 kph, which is indicative of a high mechanical coupling efficiency.

Overall, the program's feedback implied a coherent biomechanical reasoning, similar to a coach's analysis, consistently linking objective kinematic measures to subjective evaluations and producing outputs in 4-6s on a standard graphics processor. Consistent success with all bowlers and perspectives illustrated the adaptability of the program, and the explicit links between the measured quantities and textual output provided the transparency that is required for coaching use. This feedback program successfully converted motion information into useful, empirical performance feedback categories.

4.5 System Flow and End-to-End Integration

The complete pipeline converts a raw bowling video into quantitative metrics and structured feedback.

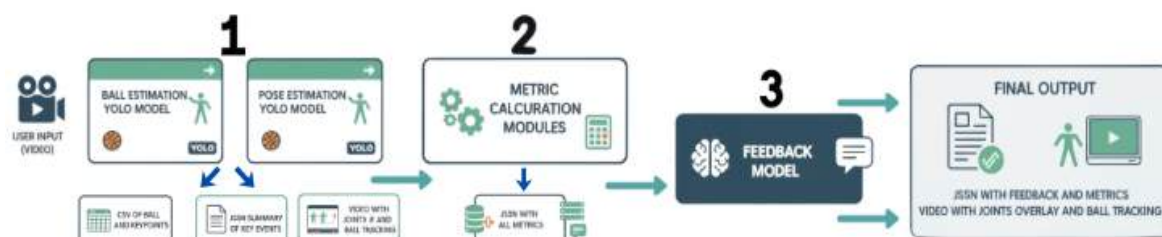


Figure 8: Complete AI Bowling Analysis Pipeline Flow

V. Discussion

Most cricket biomechanics studies have relied on expensive laboratory settings equipped with multiple-camera motion capture systems, force plates, and/or marker-based systems – these tools also create barriers to accessibility for developing programs and are unrealistic within the typical coaching process. This research addresses this gap by demonstrating how advanced biomechanical analysis is possible through the use of an unconstrained single 2-D video, leveraging computer vision and AI-based interpretation.

While most biomechanical analyses stop once the metrics have been computed, the system presented here links detection, measurement, and language-based reasoning to produce interpretable coach-ready feedback. Furthermore, the addition of a retrieval-augmented feedback model transforms raw kinematic data collected from athletes' movements into a human-readable performance assessment, associating numerical data with drillable drills and injury risk awareness. As such, this system provides a low-cost, easily accessible alternative to the motion laboratories currently used to measure the biomechanics of athlete movement, with minimal equipment requirements and limited space, while maintaining much of the current validity of cricket biomechanics.

In contrast to past biomechanical analyses that end with the collection of metrics for athletic movement patterns, this system produces interpretable, coach-ready feedback.

The addition of a retrieval-augmented feedback model allows the researcher to translate the raw kinematic measures collected from an athlete's movement into human-readable performance assessments, connecting numerical data to applicable training drills and to awareness of potential injury risk. Consequently, this study demonstrates an inexpensive, accessible alternative to the motion laboratories currently used to measure the biomechanics of athletic movement, even with limited resources (i.e., limited space and equipment), while retaining biomechanical validity.

VI. Conclusion

The research presented in this paper demonstrates that a fully automated, end-to-end framework can measure and interpret bowlers' fast bowling techniques from standard video footage. The framework uses four key components (pose estimation, ball tracking, biomechanical analysis, and a feedback generation language model) to provide users with structured data from their video recordings as well as personalized coaching advice.

Future development of the technique may be applied to many different areas and sports, including rehabilitation settings where motion is necessary but access to lab-based equipment is limited. Further work will focus on three-dimensional reconstruction using multi-camera calibration, developing real-time feedback systems, and expanding biomechanical databases to establish accurate normative ranges and improve feedback accuracy.

Ultimately, this research highlights a path toward democratizing sports biomechanics, enabling athletes and coaches anywhere to access objective, data-driven movement analysis using only accessible video technology.

Works Cited

- Baily, Lisa, et al. "Stroke Comparison between Professional Tennis Players and Amateur Players Using Advanced Computer Vision." *Proceedings of the 8th International Conference on Sport Sciences Research and Technology Support*, 2020, <https://doi.org/10.5220/0010145800440052>. Accessed 21, 2025.
- Baily, Lisa, and Jonathan Lai. "Analysis of Professional and Amateur Tennis Serves Using Computer Pose Detection." *Journal of Emerging Investigators*, emerginginvestigators.org/articles/21-168/pdf. Accessed 20 Sept. 2025.
- Bartlett, R. M., et al. "The Biomechanics of Fast Bowling in Men's Cricket: A Review." *Journal of Sports Sciences*, vol. 14, no. 5, Oct. 1996, pp. 403–24, <https://doi.org/10.1080/02640419608727727>. Accessed 5 Oct. 2025.
- C Elliott, Bruce, and Angus Burnett. "(PDF) The Biomechanics of Fast Bowling in Men's Cricket: A Review." *ResearchGate*, 9 Dec. 1995, www.researchgate.net/publication/14265183_The_Biomechanics_of_fast_bowling_in_men. Accessed 12 Oct. 2025.
- Callaghan, Samuel J., et al. "The Effects of an Eight Over Cricket Bowling Spell upon Pace Bowling Biomechanics and Performance within Different Delivery Lengths." *Sports*, vol. 7, no. 9, Aug. 2019, p. 200, <https://doi.org/10.3390/sports7090200>. Accessed 6 Oct. 2025.
- Computer vision engineer. "Train Pose Detection Yolov8 on Custom Data | Keypoint Detection | Computer Vision Tutorial." *YouTube*, 26 Apr. 2023, www.youtube.com/watch?v=gA5N54IO1ko. Accessed 28 Sept. 2025.
- Cricketeer. "Biomechanics of Fast Bowling Explained - Cricketeer." *Cricketeer - Cricket Musings Powered by Tea and TMS*, 12 Aug. 2025, cricketeer.uk/2025/08/12/biomechanics-of-fast-bowling-explained/. Accessed 10 Oct. 2025.
- OpenAI. ChatGPT, version GPT-5, OpenAI, <https://chat.openai.com/>. Used Aug.–Oct. 2025.
- DeepLearning_by_PhDScholar. "Human Pose Estimation Using OpenCV | Python | OpenPose | Stepwise Implementation for Beginners." *YouTube*, 14 Oct. 2020, www.youtube.com/watch?v=9jQGsUidKHs. Accessed 26 Sept. 2025.
- Felton, Paul, et al. "Cricket Fast Bowling: The Relationship between Range of Motion and Key Performance and Injury Technique Characteristics." *Journal of Sports Sciences*, vol. 41, no. 2, Apr. 2023, pp. 1–9, <https://doi.org/10.1080/02640414.2023.2200520>. Accessed 19 Sept. 2025.
- Glazier, Paul S., and Jonathan S. Wheat. "An Integrated Approach to the Biomechanics and Motor Control of Cricket Fast Bowling Techniques." *Sports Medicine*, vol. 44, no. 1, Sept. 2013, pp. 25–36, <https://doi.org/10.1007/s40279-013-0098-x>. Accessed 29 Sept. 2025.
- Jacobs, Jolandi, et al. "Anthropometric and Biomechanical Factors in Elite Male and Female Fast Bowlers." *Journal of Biomechanics*, vol. 170, Elsevier BV, June 2024, pp. 112157–57, <https://doi.org/10.1016/j.jbiomech.2024.112157>. Accessed 19 Oct. 2025.
- Kalman, R. E. "A New Approach to Linear Filtering and Prediction Problems." *Journal of Basic Engineering*, vol. 82, no. 1, 1960, p. 35, <https://doi.org/10.1115/1.3662552>. Accessed 1 Oct. 2025.

- Li, Wenxi, et al. "A Review of Artificial Intelligence for Sports: Technologies and Applications." *Intelligent Sports and Health*, vol. 1, no. 3, June 2025, pp. 113–26, <https://doi.org/10.1016/j.ish.2025.05.001>. Accessed 14 Oct. 2025.
- . "A Review of Artificial Intelligence for Sports: Technologies and Applications." *Intelligent Sports and Health*, vol. 1, no. 3, June 2025, pp. 113–26, <https://doi.org/10.1016/j.ish.2025.05.001>. Accessed 12 Sept. 2025.
- Loram, L. C., et al. "Determinants of Ball Release Speed in Schoolboy Fast-Medium Bowlers in Cricket." *The Journal of Sports Medicine and Physical Fitness*, vol. 45, no. 4, Dec. 2005, pp. 483–90, pubmed.ncbi.nlm.nih.gov/16446679/. Accessed 2 Oct. 2025.
- Manawadu, Kaushal. "A Biomechanical Analysis of Fast Bowling Variations." Loughborough University, June 2023, <https://doi.org/10.26174/thesis.lboro.23566362.v1>. Accessed 16 Oct. 2025.
- Marwah, Nalin. "Cricket Fast Bowling Optimization Using Machine Learning Pose Estimation Modeling." Aug. 2025, <https://doi.org/10.58445/rars.2861>. Accessed 14 Oct. 2025.
- Naik, Banoth Thulasya, et al. "A Comprehensive Review of Computer Vision in Sports: Open Issues, Future Trends and Research Directions." *Applied Sciences*, vol. 12, no. 9, Apr. 2022, p. 4429, <https://doi.org/10.3390/app12094429>. Accessed 29 Sept. 2025.
- Saeid Edriss, et al. "Commercial Vision Sensors and AI-Based Pose Estimation Frameworks for Markerless Motion Analysis in Sports and Exercises: A Mini Review." *Frontiers in Physiology*, vol. 16, Frontiers Media, Aug. 2025, <https://doi.org/10.3389/fphys.2025.1649330>. Accessed 19 Sept. 2025.
- Senington, Billy, et al. "Are Shoulder Counter Rotation and Hip Shoulder Separation Angle Representative Metrics of Three-Dimensional Spinal Kinematics in Cricket Fast Bowling?" *Journal of Sports Sciences*, vol. 36, no. 15, Dec. 2017, pp. 1763–67, <https://doi.org/10.1080/02640414.2017.1416734>. Accessed 14 Oct. 2025.
- Thiagarajan, K., et al. "Cricket Biomechanics Analysis of Skilled and Amateur Fast Bowling Techniques." *Journal of Postgraduate Medicine, Education and Research*, 2015, www.semanticscholar.org/paper/Cricket-Biomechanics-Analysis-of-Skilled-and-Fast-Thiagarajan-Parikh/bb38814cd3f27df9a8cd2c55d46b4384178db79f. Accessed 4 Oct. 2025.
- Ultralytics. "Ultralytics YOLO11 Pose Estimation Tutorial | Real-Time Object Tracking and Human Pose Detection." YouTube, 25 Oct. 2024, www.youtube.com/watch?v=AAkfToU3nAc. Accessed 28 Sept. 2025.
- Worthington, Peter J., et al. "Relationships between Fast Bowling Technique and Ball Release Speed in Cricket." *Journal of Applied Biomechanics*, vol. 29, no. 1, Feb. 2013, pp. 78–84, <https://doi.org/10.1123/jab.29.1.78>. Accessed 11 Oct. 2025.
- Ye, Pinchu, et al. IMPERIAL COLLEGE LONDON DEPARTMENT of COMPUTING 3 RD YEAR SOFTWARE ENGINEERING GROUP PROJECT Revolutionizing Tennis with Machine Learning and Computer Vision. 2019, gml16.github.io/projects/Report_LV8_Project.pdf. Accessed 15 Sept. 2025.

Losing Quechua and Oneself: The Mental Health Impacts of the Extinction of Languages

By Marco Kang Pastrana

Abstract

Every two weeks, a language disappears and with it, a culture is endangered. In this review paper, the psychological and community-level impacts of endangered language loss are identified with an emphasis placed on Quechua, which is part of the largest Indigenous language family in South America. The author of this paper combines the findings of UNESCO reports, peer-reviewed journal articles, and case studies to explore the interplay between language loss, mental health, and social cohesion. The methodology applied in this article is primarily a literature review on language endangerment and its cultural implications, thus incorporating knowledge from both psychology and linguistics. Although there are few studies on the emotional impact of Quechua language loss at the present time, the results related to other endangered languages show that losing the language leads to greater separation and loss of communication between generations, lower self-respect, and division among the people living in the affected area. At the same time, school, the internet, and revitalization efforts show promising potential for preventing language loss and restoring the identity and pride that come with language. This critique highlights the way Quechua language decline reflects broader global patterns of Indigenous exclusion, but also shows the importance of revitalization for mental health and cultural strength. Limitations of this project involve reliance on secondary sources rather than original fieldwork, and the limited studies focused specifically on Quechua that exist about this topic. The main conclusion of this review article is that mental health and cultural identity are key detrimental effects of language loss, but these effects can be ameliorated through interventions.

Introduction

Research has established an evident link between loss of Indigenous languages and adverse social and psychological outcomes. Canadian and American research describes Indigenous youth cut off from their native languages as having higher rates of depression, drug abuse, and suicide. As Hallett et al. argue in “Aboriginal language knowledge and youth suicide,” “youth suicide rates effectively dropped to zero in those few communities in which at least half the band members reported a conversational knowledge of their own Native language” (Hallett et al., 2007, p. 393). Similarly, a study in *The Lancet Psychiatry* reported that Indigenous youth with less knowledge of their native language are “six times more likely to have suicidal ideation than those with greater language knowledge” (Faruk, 2022, par. 1). The same outcomes can be observed in Latin America, where assimilation and displacement policies have dissuaded transmission of Quechua (McKenzie, 2022, p. 72).

Language loss threatens cultural diversity, collective memory, and intergenerational connectedness. Some of the consequences of language loss among speakers of Quechua are displacement and discrimination. Urban migration leads to language loss, and while Quechua is spoken at home and in community life in rural towns, the Spanish language dominates schools,

workplaces, and media in urban areas of Peru. Parents generally abandon instruction of Quechua at home so as not to inflict a disadvantage on their children. The result has been a breakdown of intergenerational transmission of language and erosion of social bonds.

Research indicates that one of the major consequences of the concealment of Quechua identity to escape stigmatization is the feeling of guilt and emotional suffering (Whalen, 2022, p. 2). This idea is supported by worldwide studies showing that the mastery of a language correlates positively with the level of one's psychological well-being. Language is the main medium through which all the Quechua tribal customs - songs, farming festivals, moral education and religious rituals - are communicated. The absence of a common language creates a gulf between people who are neighbors, and gradually the social ties are broken. Furthermore, language loss can cause the extinction of agricultural and environmental knowledge if not actively prevented.

The results of these studies point to the fact that the loss of language among speakers of Quechua and other endangered languages impacts mental health, intergenerational relationships, and community morale. The present article will therefore discuss the psychological aspects of Quechua language loss by dividing it into four subsections. It begins with a migration and displacement impact as a cause of accelerated decline of Quechua. The second subsection addresses the individual psychological conflicts of those who undergo a language decline, and the third one discusses the community and intergenerational aspects of language decline. The last part looks at language revitalization efforts that can elicit language loss prevention and restore the positive effects it has on societies. The paper concludes that language preservation is a necessity for communication as well as for preserving the identity of the people that speak it, their culture and their connectedness.

Section 1: Displacement and Language Loss

Displacement is one of the major dangers to the Quechua speaking community. Native families are frequently compelled to abandon their settlements in search of jobs or better living conditions in the urban areas. While in rural towns, Quechua is still spoken in households, markets, and during festivities, in the cities Spanish is the main language used in schools, workplaces, television, and the internet. City-raised children rarely hear Quechua outside the home. This has long-term consequences for families. If parents stop teaching Quechua at home, children grow up with limited knowledge of the heritage language. Parents abandon Quechua because they believe that using Spanish alone is the route to becoming successful in education and employment (Briceño, 2021, par. 9-10). In a high-competition job market, parents are concerned that their children will be discriminated against or disadvantaged if they are not proficient in Spanish. Fear of discrimination is also a key factor that leads to language loss. As Briceño reports about an 11-year-old girl living in Peru, “children who speak Quechua at her school get bullied, and added that parents in her village don’t want children to learn the tongue because they think it will not help children when they move to the cities for work” (Briceño, 2021, par. 7). Some even prefer not to disclose the fact that they can speak Quechua to avoid prejudice. Displacement-induced language loss is carried through generations as children watch

their parents avoid using Quechua and instead of pride, feel embarrassment, which erodes their identity and confidence.

Government policies have also contributed negatively to language loss. Peruvian schools have been encouraged to follow Spanish-only instruction for decades, at times punishing children for using Quechua in class (McKenzie, 2022, p. 73). Because Spanish is the language used in school, textbooks, the public, and the media, kids associate it with the language of modernity and education. On the other hand, Quechua is not used outside rural towns between families. This leads to children associating Quechua with the language of the past, the language they use to speak to their grandparents. As a result, generations of people have begun to distance themselves from Quechua even though it was the language of their ancestors. The same pattern has been found in other endangered languages, as in Papua New Guinea, where national language policies have brought about a decline in Indigenous languages, reducing cultural knowledge and community solidarity (Kik et al., 2021, p. 2).

Displacement causes geographical as well as emotional distance. As families migrate to the city from rural areas, grandparents get left behind in the villages. This results in difficult family communication barriers (UNESCO, 2019, par. 13). Elders can no longer impart their histories or transmit their customs, and young people miss out on inheriting centuries of wisdom built up over time that have kept communities strong. Such voids create sadness and loneliness among the elderly and detachment and confusion among the young people. Displacement breaks down the communal bonding that once held villages together. Ceremonies and rituals of traditional cultures often employ Quechua for songs, prayers, and oral accounts. In the city, these traditions are harder to maintain because individuals are dispersed, busy with work, and are subjected to Spanish as the language of choice (The Guardian, 2024, par. 3-4, 7).

With no shared environment in which to use Quechua as a language, culture declines, and people lose the feeling of community that is gained from gathering together around a shared language. Displacement, hence, changes identity, family life, and psychological well-being. For Quechua speakers, exiting their towns often means exiting the rhythm, sound, and culture of their language. Spanish becomes the common language, whereas Quechua begins to disappear. This disappearance is provoked by feelings of shame, isolation, and loss of culture, which together threaten the survival of the language and the psychological health of those who speak it.

Section 2: Mental Health Impacts of Quechua Language Loss

Studies in Papua New Guinea have shown that youth experience shame and identity dilemmas, and older generations report sorrow and isolation due to intergenerational estrangement (Kik et al, 2021, p. 4). Language loss is a social phenomenon that affects not just the way a person speaks but also the way they perceive themselves. The majority of the Quechua-speaking population considers the loss of their native language as a watershed moment in their mental state, which leads to overwhelming feelings of shame, fear, and isolation. The urban Quechua-speaking community is frequently scorned as a group lacking knowledge and is

pushed out of the mainstream (Briceño, 2021, par. 5). Such an attitude causes people to think that their language is a matter of shame and, consequently, they tend to cover up a part of their identity in order to be accepted in the society that surrounds them. Over time, it erodes self-respect and splits identity, with the negative impact being most pronounced among the youth. For example, Spanish-speaking children inevitably come to associate Quechua with ignorance and being backward, and consequently, they tend to feel ashamed of it. They might even refuse to discuss it with their parents for fear of being teased. Parents often feel guilty or mournful when their children do not want to learn and speak the language that is native to the family because they think they are losing their cultural heritage. On the other hand, the elderly feel disregarded because their grandchildren cannot communicate with them. This breakdown of communication leads to a family environment full of empathy, sadness, and, most importantly, depression.

Similar results are represented in other Indigenous research studies. Less engaged Canadian Aboriginal youth experience more pronounced suicide and mental health issues, which supports the high correlation between language retention and psychological health (Boksa, 2015, p. 365). The process of losing a language reverts back to that of a 'historical trauma', echoing the past and repercussions of stress in the community through generations. The consequences of losing a language could very much be the same as the consequences of cultural trauma (McKenzie, 2022, p. 72). Folks who are deprived of their language would also be deprived of the songs, ceremonies, and stories that used to be their source of strength. The absence of these traditions means that people cannot cope with stress and feel alienated in a society that refuses to accept them. Language is linked to well-being because it provides a sense of belonging.

When people cannot use their language, they are not only cut off from society but also from their own identity. This internal struggle results in the development of serious mental health issues that take a long time to heal. At the same time, the silence around Quechua negatively impacts the identity of the youth. Youth are particularly affected as they are undergoing a process to define their identity. If they are told as kids that Quechua is something to be ashamed of or is backward, they tend to carry that stigma into their adult lives. This often leads them to sever their ties with their culture so as not to be shunned, yet, this also leads to a sense of loss. Instead of being able to identify with pride, they will have confusion about their social positioning. The extinction of Quechua is not merely about the loss of a language; it is about the loss of the self-assuredness, the pride, and the identity that the speakers were bound to feel. In the absence of Quechua, a large part of the population feels invisible in their own land, causing detrimental psychological consequences.

Section 3: Shattered Community Ties and Social Isolation

Language loss affects the cohesion of a community. In Quechua-speaking communities, language is a vehicle for community ceremonies, decision-making, and community existence. The loss of Quechua affects the sense of belonging to a community that comes from having a shared language. Community traditions like “fiestas,” markets, and oral history transmission

events are lost in cities because of the pervasive influence of Spanish (Briceño, 2021, par. 21-23). At the national and global levels, endangered languages are inextricably connected with the loss of ecological and medicinal knowledge, showing that when languages vanish, the health and expertise of communities also deteriorate (Cámara-Leret & Bascompte, 2021, p. 1). Cámara-Leret and Bascompte found that more than 75 percent of medicinal knowledge in New Guinea was “linguistically unique” and therefore permanently lost when that language disappeared (Cámara-Leret & Bascompte, 2021, p. 1).

One of the most apparent effects is the loss of ties between generations. Older individuals who only speak Quechua cannot communicate with grandchildren who only speak Spanish. This results in elders feeling alienated from younger generations. Failure to pass language impacts their capacity to transmit their values, culture, and shared memories and leads to their social isolation (Khawaja, 2021, p. 5). At a general level, Quechua decline contributes to the erosion of Indigenous rights. With the decrease in the number of speakers, it becomes harder for smaller communities and groups to make political claims or request cultural protection. Endangered languages will contribute to the loss of global cultural diversity, and as UNESCO has repeatedly warned: “at least 40% of the world’s languages are endangered, with one language disappearing every two weeks” and this loss continues to erode cultural diversity worldwide (UNESCO, 2025, par. 4). For Quechua, this will mean the loss of a key cultural heritage from the Andes communities in Peru.

Section 4: Revitalization and the Value of Language Preservation

There are rays of hope amidst these challenges. Revitalization efforts across the Andes prove that Quechua can survive and prosper if schools, governments, and communities make efforts to keep it alive. Language programs in schools are opening doors to learning Quechua for kids in school. For example, in Ayacucho, Peru, regional authorities introduced a bilingual education program in 2018 that now reaches more than 35,000 students. Across the country, the impact is even broader: currently, “more than 1.2 million children receive bilingual education in Spanish and Indigenous languages, and the government aims to increase the number of bilingual teachers in public schools from 54,000 to around 60,000 next year” (Briceño, 2021, par. 24). These initiatives are preventing the negative effects of language loss, and those who were once ashamed of their language are now recovering the pride in their identity through Quechua. When kids express themselves in Quechua without hesitation, their parents and the whole community gradually regain their self-assurance as well.

Local communities have been key to the success of these initiatives and activities. Daily life is being infused with Quechua through festivals, workshops, and storytelling sessions. These activities, besides preserving the language, also revitalize the bonds between the old, the middle-aged, and the young. This process of becoming one again through language is supported by McKenzie’s claim that “we must also provide supportive space for all speakers to experience healing as they engage in our languages” (p. 74). This healing process is characterized by the building of bridges and the erasing of isolation caused by the vanishing of the language

(UNESCO, 2019, par. 16-18). Language revitalization indicates the survival of the language but also mental recovery; hence, “in this sense our languages can literally heal us” (McKenzie, 2022, p. 75). Communicating in Quechua means that people are not only reclaiming their pride but also their identity. Communities that support their language enjoy more trust, more social cohesion, and more confidence to confront discrimination. Indigenous language revitalization also has measurable health benefits as it leads to less depression and stronger social health.

Government recognition is very important as well. Whenever Quechua appears on public signs, official documents, and mass media, it becomes more legitimate. It is a very clear sign that Quechua is not a thing of the past but rather a language for today and tomorrow. Official status is the life-saver for the long-term existence of a language because it signifies that Quechua is not only supported by the communities but also maintained at the national and international levels. Hence, the process of revitalization is cultural as well as psychological. It brings people back to their true selves, reminding them that their language is a power and not a liability. For families, it reaffirms and strengthens intergenerational relationships. For communities, it opens new avenues for closeness. And for individuals, it brings back the feeling of pride.

Conclusion

The deterioration of the Quechua language concerns not only the loss of a language but also the loss of people's identity, mental health, and culture. The present review article has shown how the decline of the Quechua language can be associated with seclusion, psychological problems, and the collapse of social connections. A person who speaks an indigenous language and loses it can suffer from feelings of disgrace, alienation, and low-self esteem, while the society that once spoke the language loses one of its major intergenerational links and cultural inheritance mechanisms. Efforts made by schools, festivals, and celebrations along with international attention have started to show their effects on the sustainability and continuity of the Quechua language (Briceño, 2021, par. 24). Other endangered languages have been helped through similar initiatives and these revitalization efforts have even been able to restore ecological learning and healing practices (Cámara-Leret & Bascompte, 2021, p. 1). As this review shows, language conservation serves two purposes: the reduction of the negative psychological effects of language loss and the preservation of cultural bonds and traditions. The act of learning, teaching, and speaking Quechua helps the individual and the family to regain their pride, revive their intergenerational connections, and resist the effects of past discrimination. Quechua's survival is hinged on the long-term commitment to revitalization and acknowledgement. The process of language conservation is not only about saving the vocabulary but about saving the humans' mental and moral health and their social strength and cohesion. Every kid who is being taught the Quechua language, every elderly person who is telling a story, and every community that is still practicing its language are all taking part in an important process of language survival and mental healing. When it happens, language transforms from being merely a means of communication into a way to claim one's dignity, to achieve cultural continuity, and to bring hope for the future.

Works Cited

- Kik, A., et al. "Language and Ethnobiological Skills Decline Precipitously in Papua New Guinea, the World's Most Linguistically Diverse Nation." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 118, no. 22, 2021, e2100096118. <https://doi.org/10.1073/pnas.2100096118>
- Faruk, Md Omar, and Simon Rosenbaum. "The Mental Health Consequences of Indigenous Language Loss." *The Lancet Psychiatry*, vol. 9, no. 10, Oct. 2022, p. e46. doi:10.1016/S2215-0366(22)00300-5. [https://www.thelancet.com/journals/lanpsy/article/PIIS2215-0366\(22\)00300-5/fulltext?utm_source=chatgpt.com](https://www.thelancet.com/journals/lanpsy/article/PIIS2215-0366(22)00300-5/fulltext?utm_source=chatgpt.com)
- Cámara-Leret, Rodrigo, and Jordi Bascompte. "Language Extinction Triggers the Loss of Unique Medicinal Knowledge." *Proceedings of the National Academy of Sciences*, vol. 118, no. 24, 2021, e2103683118. *Proceedings of the National Academy of Sciences*, <https://doi.org/10.1073/pnas.2103683118>.
- Hallett, Darcy, Michael J. Chandler, and Christopher E. Lalonde. "Aboriginal Language Knowledge and Youth Suicide." *Cognitive Development*, vol. 22, no. 3, 2007, pp. 392–401. ScienceDirect, <https://doi.org/10.1016/j.cogdev.2007.02.001>.
- Holmes, Paul. "Louisiana School Brings Cajun Language Back to Life." *The Guardian*, 14 Feb. 2024, <https://www.theguardian.com/us-news/2024/feb/14/louisiana-school-indigenous-cajun-languages-community-impact>.
- Kirmayer, Laurence J., Joseph P. Gone, and Joshua Moses. "Rethinking Historical Trauma." *Transcultural Psychiatry*, vol. 51, no. 3, 2014, pp. 299–319. PubMed, <https://pubmed.ncbi.nlm.nih.gov/31618944/>.
- McKenzie, Keren. "Addressing Historical Trauma and Healing in Indigenous Language Cultivation and Revitalization." *Annual Review of Applied Linguistics*, vol. 42, 2022, pp. 70–85. Cambridge University Press, <https://doi.org/10.1017/S0267190522000050>.
- "Indigenous Languages: Knowledge and Hope." UNESCO Courier, 2019, <https://courier.unesco.org/en/articles/indigenous-languages-knowledge-and-hope>.
- "Cutting Edge: Indigenous Languages as Gateways to the World's Cultural Diversity." UNESCO, 2022, <https://www.unesco.org/en/articles/cutting-edge-indigenous-languages-gateways-worlds-cultural-diversity>.
- "Weather, Magic, and Wind Lore: The Push to Preserve Ancient Knowledge in Vanuatu." *The Guardian*, Sept. 2024, <https://www.theguardian.com/world/article/2024/sep/07/weather-magic-and-wind-lore-the-push-to-preserve-ancient-knowledge-in-vanuatu>.
- Whalen, Douglas H., et al. "Health Effects of Indigenous Language Use and Revitalization: A Realist Review." *International Journal for Equity in Health*, vol. 21, 2022, article 169. SpringerOpen, <https://doi.org/10.1186/s12939-022-01782-6>.

- Briceño, Franklin. "Quechua Endures in Peru despite Centuries of Discrimination." AP News, 2021, <https://apnews.com/article/peru-south-america-language-e9fa5f4ddd5432d1b2e203dccc05c698>.
- Boksa, Patricia, Ridha Joobar, and Laurence J. Kirmayer. "Mental Wellness in Canada's Aboriginal Communities: Striving toward Reconciliation." *Journal of Psychiatry & Neuroscience*, vol. 40, no. 6, 2015, pp. 363–65. PubMed Central, <https://doi.org/10.1503/jpn.150309>.
- Masud Khawaja et al. "Consequences and Remedies of Indigenous Language Loss in Canada." *Societies* (2021). <https://www.mdpi.com/2075-4698/11/3/89>
- UNESCO. "Languages Are Not Just Words—They're Worlds: UNESCO Bangkok Celebrates 25 Years of International Mother Language Day." UNESCO, 25 Feb. 2025, www.unesco.org/en/articles/languages-are-not-just-words-theyre-worlds-unesco-bangkok-celebrates-25-years-international-mother.

Why Does Stirring Hot Tea Make It Cool Faster? — A Geometric Study of Surface Area

By Jimin Lee

Abstract

The study investigates why stirring hot tea makes it cool faster, focusing on the surface area of the liquid exposed to air. When tea is stirred, a vortex forms such that the surface forms a cone or a truncated cone. Using geometric assumptions, the study derives the formulas for the lateral surface area of cones and frustums, and applies them to model the surface of the stirred tea. By comparing the resulting surface area with the flat surface when the tea is not stirred, the study thereby estimates the increase in the surface area in contact with air. This provides a mathematical explanation of the intuitive observation that stirring accelerates the cooling process of liquids.

Keywords

Stirring and Cooling, Surface Area Analysis, Geometric Modeling

1. Introduction

People generally tend to cool down hot tea before drinking, with one of the most common methods being stirring the hot tea with a teaspoon. When a cup of tea is stirred, a vortex forms on the surface, causing the centre of the surface to sink and creating a funnel-like shape. At this moment, the actual contact area between the surface of the tea and the air increases; as the contact area increases, heat is able to escape more easily, leading to the tea cooling faster.

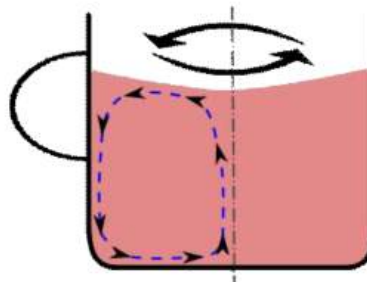


Figure 1: Schematic diagram of stirred tea [1].

The study starts with a simple idea, and focuses on how much of the change in the surface area in contact with air can be caused by stirring. The study will compute the relation between the surface shape created by stirring and the resulting area under a few assumptions. The main question is as follows: if we increase the stirring speed, how does the surface of the tea change, and how much does the contact area with air grow? Is that increase large enough to lead to twofold or threefold faster cooling?

In section 2, the study states the assumption in the vortex shape that appears when a cup of tea is stirred. In short, we assume the cup to be a cylinder, and the surface of the vortex can be approximated by a part of the lateral surface of a cone. In section 3, the study presents a formula

for the lateral area of the cone in terms of its radius and height. In section 4, based on the formula derived in section 3, the study computes the total contact area in a given situation. Finally, in section 5, the conclusion of the study based on the previous analysis is presented.

2. Assumptions on the Vortex Shape

The study utilizes assumptions to model the surface of the vortex created after stirring tea. First, the cup holding the tea is assumed to be a right circular cylinder. As tea is stirred, the centre of the cylinder moves downward and the outer region near the edge of the cup rises. To avoid overflow in the model, the study will assume that the cup extends upward like an infinitely tall cylinder.

The simplest way to model the surface of the stirred tea is by a cone. Thus, we assume that the upper surface formed by stirring can be approximated by or by a part of a cone. In cases without stirring, slow stirring, and fast stirring, the shape of the vortex can be modeled as follows:

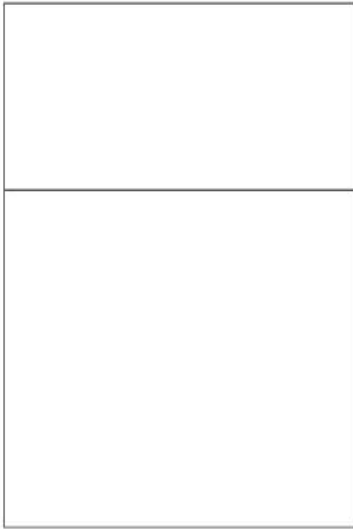


Figure 2: No Vortex Case

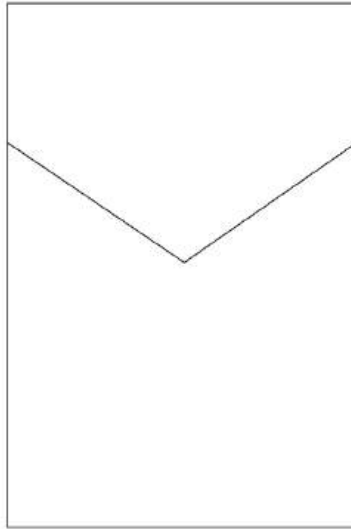


Figure 3: Small Vortex Case

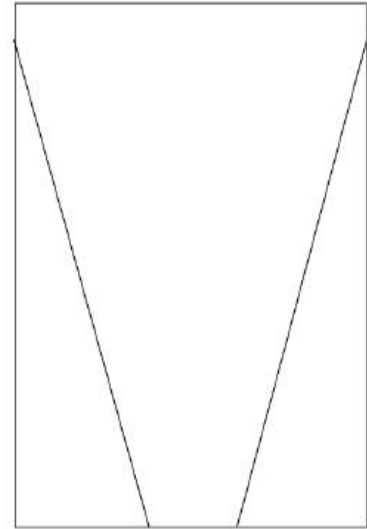


Figure 4: Strong Vortex Case

The surface of the tea is flat when the tea is not being stirred (Figure 2). With slow stirring, the center sinks slightly and the shape resembles a shallow cone (Figure 3). When the stirring becomes faster, the center of the vortex deepens until it reaches the bottom of the cup (Figure 4). At this point the surface is no longer a sharp cone but rather a frustum; the centre of the cup is exposed as a small circular region, and the liquid surface rises outward in the form of a truncated cone.

Since the analysis requires mathematically modeling the lateral surface areas of cones and frustums, the next section will derive the above formulas given the radius and height.

3. Formulas of the Lateral Surface Areas of Cones and Frustums

Let a cone have a base radius $r > 0$ and height $h > 0$. Denote by l the slant height of the cone from the apex to the rim. This cone can be illustrated in the following structure.

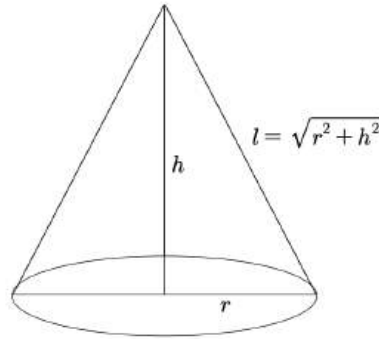


Figure 5: Geometry of a cone with base radius r , height h , and slant height l .

From the right triangle formed by the height, the radius, and the slant height, the Pythagorean theorem gives

$$l = \sqrt{r^2 + h^2}$$

Based on the cone above, the following net can be constructed.

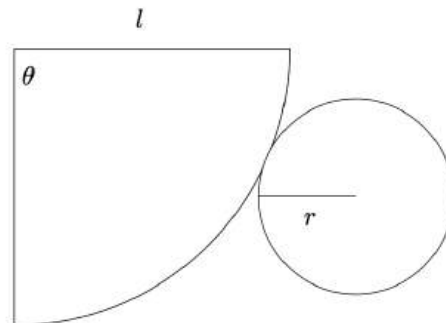


Figure 6: The net of the cone

The lateral surface of the cone yields a sector with radius l and the central angle θ . the arc length of this sector is identical to the cone's base:

$$l\theta = 2\pi r.$$

The area of the sector can be obtained as follows:

$$A = \frac{l^2\theta}{2} = \frac{(l\theta)l}{2} = \pi rl$$

Substituting $l = \sqrt{r^2 + h^2}$ yields

$$A(r, h) = \pi r \sqrt{r^2 + h^2}.$$

Here, the lateral surface area of a cone as a function of base radius and height is denoted as $A(r, h)$.

This formula can be applied for a given frustum. Denote the top and bottom radii of the frustum as r_1 and r_2 , respectively, and height as h . If $r_1 = r_2 = r$ such that the solid is a cylinder, the area of the lateral surface is $2\pi rh$.

If $r_1 < r_2$, the equation for the lateral surface of the cone can be used to calculate the lateral surface area of the frustum. The cross section of the frustum and its central axis can be demonstrated by the following figure:

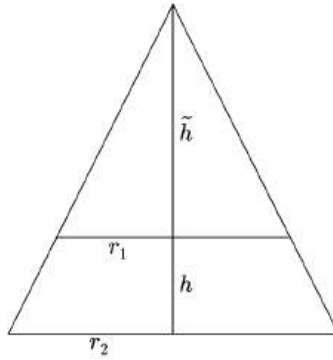


Figure 7: Cross section of a frustum with the central axis

The cross section of the frustum containing the central axis forms an isosceles trapezoid, from which the two side segments can be extended to form a triangle. Rotating this shape about its central axis yields two cones, one smaller than the other. By subtracting the lateral surface area of the smaller cone from the larger cone, the lateral surface area of the frustum can be obtained. Therefore, the surface area of the frustum can be expressed as

$$A(r_2, \bar{h} + h) - A(r_1, \bar{h}) = \pi r_2 \sqrt{r_2^2 + (\bar{h} + h)^2} - \pi r_1 \sqrt{r_1^2 + \bar{h}^2}.$$

From $r_1 : \bar{h} = r_2 : \bar{h} + h$, the equation can be rearranged as

$$r_2 \bar{h} = r_1 \bar{h} + r_1 h,$$

which implies

$$\bar{h} = \frac{r_1 h}{r_2 - r_1} \text{ and } \bar{h} + h = \frac{r_2 h}{r_2 - r_1}.$$

Finally, the above result can be substituted to the area formula, which yields

$$\begin{aligned}
A(r_2, \bar{h} + h) - A(r_1, \bar{h}) &= \pi r_2 \sqrt{r_2^2 + \left(\frac{r_2 h}{r_2 - r_1}\right)^2} - \pi r_1 \sqrt{r_1^2 + \left(\frac{r_1 h}{r_2 - r_1}\right)^2} \\
&= \frac{\pi(r_2^2 - r_1^2)}{r_2 - r_1} \sqrt{(r_2 - r_1)^2 + h^2} \\
&= \pi(r_1 + r_2) \sqrt{(r_2 - r_1)^2 + h^2}
\end{aligned}$$

4. Application to the Surface of a Stirred Tea

Based on the previous sections, this section determines the increase in the area of contact with air for each assumed shape when a cup of tea is stirred compared to when it is not stirred.

Let the radius of the cup be R , and the volume of tea inside the cup be V . For this case, the initial height of the tea can be set to $H = \frac{V}{\pi R^2}$. In addition, denote the angle of the cone that characterizes the slope of the stirred surface as α ($0 < \alpha \leq \frac{\pi}{2}$), as illustrated in the diagram below. For a steady state without a vortex, $\alpha = \frac{\pi}{2}$.

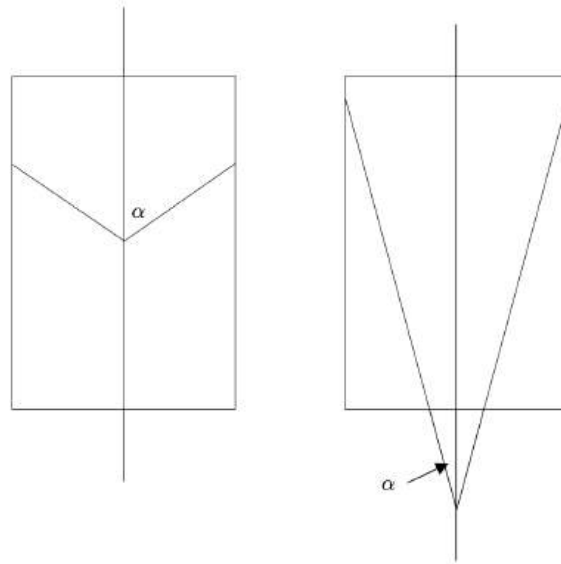


Figure 8: Geometric definition of α . α is relatively large for a slow vortex, and relatively small for a large vortex.

Consider a vortex with the vertex of the cone above the bottom of the cup as a slow vortex, and a vortex with the vertex of the cone below the bottom of the cup as a fast vortex. The expression for α in the boundary case can be found as follows:

$$\pi R^2 (R \cot \alpha) - \frac{1}{3} \pi R^2 (R \cot \alpha) = \pi R^2 H.$$

Which can be simplified into

$$\frac{2}{3} (R \cot \alpha) = H$$

and implies that $\tan \alpha = \frac{2R}{3H}$. Therefore, $\alpha = \arctan(\frac{2R}{3H})$ is the angle that divides the fast and slow vortex.

The two cases will be examined separately.

4.1. The Slow Vortex: $\arctan(\frac{2R}{3H}) \leq \alpha \leq \frac{\pi}{2}$.

For this case, the height of the cone is simply $R \cot \alpha$. Therefore, the formula gives

$$A(R, R \cot \alpha) = \pi R \sqrt{R^2 + R^2 \cot^2 \alpha} = \pi R^2 \csc \alpha = \frac{\pi R^2}{\sin \alpha}.$$

as $\csc^2 \alpha = 1 + \cot^2 \alpha$. Since the range of α is $\arctan(\frac{2R}{3H}) \leq \alpha \leq \frac{\pi}{2}$, it can be determined that

$$\frac{2R}{\sqrt{9H^2 + 4R^2}} \leq \sin \alpha \leq 1.$$

Therefore, the maximum lateral surface area of the cone for this case is

$$\frac{\pi R^2}{\sin \alpha} \leq \frac{1}{2} \pi R \sqrt{9H^2 + 4R^2}.$$

4.2. The Fast Vortex: $0 \leq \alpha \leq \arctan(\frac{2R}{3H})$.

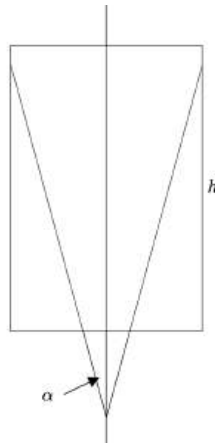


Figure 9: Relationship between h and α for a fast vortex

Here, r_2 can be determined as R . Since the total volume of the tea is $\pi R^2 H$, the following equation can be obtained:

$$\pi R^2 h - \frac{1}{3} \pi R^2 (R \cot \alpha) + \frac{1}{3} \pi (R - h \tan \alpha)^2 (R \cot \alpha - h) = \pi R^2 H.$$

This can be simplified by dividing both sides by $\pi R^2 H$ as follows:

$$\frac{h}{H} - \frac{R \cot \alpha}{3H} + \frac{R \cot \alpha}{3H} \left(1 - \frac{h}{R} \tan \alpha\right)^3 = \frac{1}{3H} \left(\frac{3h^2 \tan \alpha}{R^2} - \frac{h^3 \tan^2 \alpha}{R^2}\right) = 1.$$

Since this is a cubic equation, the value of h cannot be found directly but it is possible to guarantee that there is a unique solution h for $0 < \alpha \leq \arctan(\frac{2R}{3H})$ geometrically. Overall, after finding the estimated value of h numerically, the values $r_1 = R$, h , and $r_2 = R - h \tan \alpha$ can be substituted to determine the new lateral surface area of the frustum.

5. Conclusion

This paper models the surface of a stirred tea as either a cone or a frustum, and calculates the corresponding lateral surface areas exposed to air. By comparing these areas with the case without stirring, the study quantified how stirring enlarges the contact surface with air. The analysis shows that the vortex formed from stirring indeed increases the exposed area, supporting the idea that stirring accelerates cooling. While the exact ratio depends on the steepness and the depth of the vortex, the geometric framework demonstrates clearly that a stronger stirring produces a larger exposed surface. Thus, the simple act of stirring a cup of tea can be understood as a geometric mechanism for enhancing heat loss.

Works Cited

- [1] Apoc2400, Tea leaf Paradox Illustration.svg, Wikimedia Commons, CC BY-SA 3.0.
- [2] Simmons, G. F. Precalculus Mathematics in a Nutshell. Waveland Press, 1997.
- [3] Muller, R. A. Physics for Future Presidents: The Science Behind the Headlines. W. W. Norton, 2008.

Artificial Intelligence in Alzheimer's Disease GWAS By Joy Lukama

Abstract

Alzheimer's disease (AD) is a progressive brain disorder that deteriorates one's cognitive functions. AD is caused by the accumulation of abnormal proteins called amyloid plaques and tau tangles within the brain. Currently, there is no cure for AD. In 2005, researchers identified genome-wide association studies, or GWAS, to be one of the most effective methods of locating specific genomic regions associated with diseases. Recent advances in artificial intelligence (AI) and machine learning (ML) have the potential to significantly reduce the time and complexity involved in conducting genome-wide association studies (GWAS). This narrative literature review explores how AI and ML are being integrated into GWAS to accelerate the discovery of genetic risk factors for Alzheimer's disease and improve future diagnosis and treatment strategies.

Method

This narrative literature review examines various publications from 2015-2025 in order to explore the integration of artificial intelligence within genome-wide association studies and Alzheimer's disease. Literature research was performed using databases such as PubMed and Google Scholar from September to October of 2025. Core search terms included combinations of "Alzheimer's disease," "GWAS," "genome-wide association," "artificial intelligence," "machine learning," "deep learning," and "SNP prediction." Additional sources were obtained via the MedRxiv preprint server, National Institute of Health (NIH) website, and direct journal access. Inclusion criteria prioritized empirical studies and review articles covering AI/ML methods in AD genetic research, published in peer-reviewed journals or reputable preprint servers. Studies included focus primarily on human data; however, computational method development was also included. Authoritative clinical and educational sources were incorporated to ensure understanding of the foundations of AD pathology. This is a narrative review designed to synthesize key developments and identify emerging trends rather than a systematic review with exhaustive search protocols.

Introduction

Alzheimer's disease, or AD, is a cognitive disorder in which the patient experiences cognitive deterioration, memory loss, and confusion. AD stems primarily from amyloid plaques and neurofibrillary tangles that form in the brain. Amyloid beta, a protein fragment derived from amyloid precursor protein (APP), accumulates between neurons and prevents neuronal communication. Furthermore, the formation of tau tangles, another hallmark of the disease, is another primary cause of AD. The amyloid plaques, in combination with tau tangles, result in the deterioration of the patient's cognitive function. Several factors increase the risk of AD, including age, genetics, lifestyle, and environmental exposures. Patients suffering from AD tend to experience the progressive loss of physical and mental abilities, such as memory and

movement. This progressive loss of function can lead to severe infections, malnutrition, and several other behavioral and psychological problems. Because these complications can be fatal, early detection of AD is imperative. Identifying AD can be done by examining the patient's genes for risk loci, regions of the genome that carry genetic variants that code for a specific disease.

Genome-wide association studies (GWAS) are used to identify genetic regions, or risk loci, associated with disease. Using GWAS, scientists have been able to more closely examine genomes for particular genetic variants, otherwise known as single-nucleotide polymorphisms (SNPs), that indicate mutations. These genome-wide association studies are very complex and can take anywhere from a month to a year to complete. Despite their complexity, GWAS are essential for understanding the genetic basis of diseases like AD. GWAS has helped researchers find genetic variants linked to diseases, even in non-coding regions, which could eventually lead to new gene-based treatments. GWAS have been an imperative tool within the genetics industry and will only become more useful as our technologies advance. As of 2023, GWAS have identified over 400,000 SNP-trait associations from more than 45,000 individual GWAS (Sollis et al.). However, this amount is nowhere near the potential number of genetic associations GWAS could identify in the coming years with the use of new technology.

GWAS discovery rates have grown rapidly over the past decade and are expected to continue expanding as sample sizes increase and computational methods improve. This rapid increase in identification is revolutionary considering that the GWAS Catalog contained significantly fewer associations in its earlier years (Sollis et al.). The increase in identifications is likely due to the fact that many of the constraints that were faced regarding GWAS in the past no longer play a pivotal role in the process today. For example, genotyping costs have decreased dramatically over time, and modern GWAS can now be conducted with large sample sizes at significantly lower costs than in the early 2000s (Uffelmann et al.). Time has also played a large role in the process of GWAS. Advances in computational infrastructure and analysis pipelines have reduced the time required to complete GWAS analyses substantially compared to the early days of technology (Uffelmann et al.).

While constraints such as time and cost have been alleviated over time, there are still improvements that can be made to the GWAS process with the use of AI and ML technology. AI and ML technology have contributed greatly to the biotechnological field. For example, because the GWAS data primarily comes from those of European descent, the data generated by GWAS may reflect population bias, which would prevent scientists from being able to take into account various ethnic-specific risk loci within the genomes. However, various AI models can be trained with ethnically diverse datasets and provide more generalized data. This approach would be especially beneficial as it will allow scientists to identify any ethnic related factors that pertain to AD.

1. Alzheimer's Disease

1.1 Pathology of Alzheimer's Disease

The hallmarks of Alzheimer's Disease (AD) are amyloid plaques and tau tangles, which interfere with normal brain function. Under normal brain conditions, amyloid precursor proteins (APP) aid in neural development, synaptic formation and repair, as well as neuroplasticity. However, when APP is improperly fragmented, it produces beta-amyloid fragments, which, when accumulated, create plaques that settle between nerve cells, causing a disruption in cell communication as well as inflammation of brain tissue. Tau is another protein that regulates the body's neuronal microtubules. Tau is an imperative protein within the body; however, in AD, a chemical alteration within the tau proteins causes them to bind to each other instead of their microtubules. The accumulation of tau causes it to form threads that tangle amongst each other and disrupt the transport system of the neurons, which allows for the removal of neuronal waste. Without functioning transport and waste removal systems, the neurons find themselves overwhelmed with the amount of waste, and this causes cell death within the brain. As a result of the amyloid plaques in combination with tau tangles, the patient's cognitive function begins to slowly deteriorate.

1.2 Genetics of Alzheimer's Disease

Alzheimer's Disease (AD) can occur in one of two forms: sporadic or familial. Sporadic AD is the most common form of AD, seen in approximately 95% of all AD patients (Bird). Sporadic AD stems from a combination of risk factors, including genetics, lifestyle, and environment. The presence of the APOE ϵ 4 allele is a major risk factor for sporadic AD; however, its presence does not guarantee the presence of AD. Typically, those who develop sporadic AD encounter symptoms in their mid-60s or later ("What Are the Signs"). In contrast, familial AD is an inherited form of the disease. Familial AD can typically be identified in a patient who carries mutations in specific genes such as APP, PSEN1, and PSEN2. Because familial AD is inherited, patients typically exhibit symptoms earlier than those with sporadic AD, generally in their 30s or 40s, although this early onset can occur as late as the early 60s ("Early-Onset Alzheimer's Disease"). However, familial AD is rare and can be observed in less than 1% of all AD cases (Bird).

GWAS compares the genomes of various individuals and identifies genetic variants, or risk loci, associated with disease. Once a risk locus is identified, scientists will take a closer look at any single-nucleotide polymorphisms (SNPs) within that region to identify which specific genetic variants are more common in individuals with the disease. It is important to note that risk loci are not the direct cause of disease; rather, they are regions within the genome with genetic susceptibility. The more risk loci that are present within a genome, the more likely an individual may develop the disease; however, it does not guarantee it. That being said, GWAS is not effective at detecting rare mutations that cause familial AD, as the genetic variants occur too infrequently in the general population to be identified with GWAS. In contrast, GWAS can be particularly effective in studying sporadic AD, as there are large enough sample sizes to detect meaningful genetic associations.

1.3 GWAS in Alzheimer's Disease Research

GWAS (Genome-Wide Association Studies) scans the genomes of various individuals for recurring genetic variants, specifically SNPs, that indicate the possibility of disease. The process begins with sample collection, in which scientists gather thousands of samples from varying individuals who have been affected by the disease (cases) and have not been affected by the disease (controls). GWAS then scans each sample for single nucleotide changes located within regions of the genome associated with increased risk using a method called genotyping. Scientists then conduct a statistical analysis in which they compare the frequency of specific SNPs in the control and case groups to determine which variants are significantly associated with the disease. Scientists ensure to take into account various factors such as race, gender, and age to avoid population stratification bias. GWAS is only able to identify occurrences of genetic variants with great statistical significance, which is why GWAS is ineffective in detecting rare mutations, such as familial AD.

The APOE ϵ 4 allele is the most common genetic risk factor of sporadic AD. The presence of the APOE ϵ 4 allele renders patients with a 2-3 times higher risk of AD when carrying one copy (Yamazaki et al.). The APOE ϵ 4 allele was identified as a risk indicator of AD before the development of GWAS, which demonstrates the significance of its impact on AD patients. Aside from APOE ϵ 4, GWAS has identified various other genetic risk loci for AD, such as CLU, PICALM, BIN1, CR1, TREM2, and ABCA7, many of which are involved in our immune response, lipid metabolism, and neuronal function (Wightman). For example, BIN1 and PICALM regulate synaptic function and neuronal communication—processes known to be disrupted in Alzheimer's disease. GWAS findings allow for early disease detection and new targets for drug development.

Despite many advances with GWAS, it still faces many challenges. Small effect sizes make it difficult for scientists to know which SNPs meaningfully contribute to disease risk. This means that scientists cannot rely on GWAS findings alone when developing clinical predictions or personalized medicine. Additionally, conducting GWAS requires very large datasets, which are pooled from consortia such as the International Genomics of Alzheimer's Project (IGAP).

Collecting and processing these large datasets is highly resource-intensive, and because of that, it often excludes underrepresented populations. Population bias is also a primary constraint, as many of the samples collected are from those of European descent, which leaves out genetic risk factors that may be tied to race or ethnicity. Because many of the risk loci are located in non-coding regions of the DNA, it is unclear how many of the genetic variants directly influence the disease. This lack of functional understanding makes it difficult for scientists to develop drugs or interventions to combat the effects of the variants. These challenges highlight the need for more advanced approaches, such as artificial intelligence, to analyze the genetic data presented by GWAS.

2. Artificial Intelligence

2.1 AI and ML Fundamentals

Artificial Intelligence (AI) is a field of computer science that allows computers to perform tasks that typically require human intelligence to complete. This can include recognizing patterns, solving problems, or making decisions. Machine learning (ML) is a subset of AI focusing on teaching computers how to learn from data without necessarily being programmed to do so. This learning aspect allows for ML algorithms to improve over time. Deep Learning (DL) is a specialized form of ML that allows computers to learn using layered structures called neural networks, which allow computers to perform more complex analysis. DL allows AI to perform far more intricate tasks, such as image recognition, speech processing, and analyzing large genetic datasets. AI and ML have been revolutionizing biomedical research. These tools have played a major part in the way scientists understand and treat diseases. From analyzing patient data for early disease detection to predicting disease outbreaks, AI and ML will only become more useful in this field in the coming years. In genomics, AI's ability to analyze large datasets and recognize patterns has made it significantly easier for researchers to identify genetic variants in diseases like Alzheimer's disease through GWAS.

2.2 AI in Genomics

One human genome contains over three billion base pairs and millions of single-nucleotide polymorphisms (SNPs) (Levy). In genome-wide association studies (GWAS), data comes from tens or even hundreds of thousands of individuals. The massive scale of genomic datasets makes it difficult and time-consuming for researchers to analyze them using traditional statistical approaches, which often examine one SNP at a time. This is where artificial intelligence (AI) becomes an asset to the field. Pattern recognition refers to AI's ability to detect meaningful structures and relationships within large datasets. While traditional statistical methods may miss the subtle, more complex patterns, AI is able to identify them as it is designed to analyze thousands of genetic variables at once. Deep learning (DL) models, which use neural networks, can be particularly effective in analyzing the complexity of this enormous amount of genomic data. This capability is particularly useful within Alzheimer's disease (AD) research, as AI can find patterns that associate various SNPs with increased disease risk. Additionally, predictive models can be developed for various diseases. Predictive models use AI to forecast how likely it is that a patient is going to develop Alzheimer's disease. Predictive models supersede traditional models in accuracy because they use clinical and biological data to analyze complex interactions between thousands of genetic variants to predict a patient's risk for disease. A few machine learning (ML) techniques that are particularly effective at uncovering predictive patterns are random forests, support vector machines (SVMs), and deep learning neural networks, as they are capable of handling large-scale data. GWAS focuses on identifying individual risk loci; however, predictive modeling is able to examine the complex interactions between genetic variants and how they influence each other in order to improve the accuracy of risk prediction and show scientists which risk variants to prioritize for further research. Today, predictive models have been used to stratify patients by risk, enabling early intervention or

targeted therapeutic strategies. Predictive modeling advancements showcase the boundless potential of AI-driven genetic research in Alzheimer's Disease. Traditional GWAS is limited in its ability as it is unable to cope with the growing complexity and scale of genomic data. Artificial intelligence will serve as a powerful asset in allowing for larger datasets to be interpreted at a faster rate of time. Because traditional GWAS relies on linear models, it can only scan one SNP at a time, making it difficult to examine the interactions between genetic variants. As datasets become increasingly large in size and dimensionality, traditional statistical methods struggle to scale and maintain predictive accuracy. In contrast, because of AI's ability to conduct multivariable analysis, it is able to process various SNPs at once and detect non-linear, higher-order interactions.

Moreover, AI has the ability to scale more efficiently with larger datasets than traditional statistical methods while still maintaining predictive efficiency and accuracy, as it is highly adaptable. AI is also able to incorporate a more diverse range of data types, such as proteomics, neuroimaging, and clinical records, creating more holistic and accurate predictive frameworks. Deep learning is particularly useful in this, as it is compatible with the high-dimensional, multi-layered datasets of genomics. A 2021 study used a neural network model to predict Alzheimer's-related SNPs from whole-genome data and achieved significantly higher prediction accuracy compared to traditional logistic regression models (Monk et al.). These advantages showcase the pivotal role AI will play within genetic research in Alzheimer's disease.

2.3 Limitations of AI

While artificial intelligence poses impactful solutions to the limitations of traditional GWAS and genomics methods, AI brings its own set of limitations that must be carefully addressed to ensure reliable and equitable clinical application. Machine learning's ability to train and learn based on the data provided plays a large role in AI's pattern recognition efficiency. However, due to the fact that ML learns solely based on the provided datasets, it may be subject to bias, as GWAS data may not include that of non-European populations. This population bias risks producing models that perform poorly in underrepresented populations, potentially missing ethnicity-specific risk variants or generating inaccurate predictions for non-European individuals (Duncan et al.). To mitigate this, researchers can employ ancestry-stratified training, where separate models are developed for different ancestral groups, or utilize transfer learning approaches to adapt models trained on larger datasets to work effectively in smaller, diverse cohorts (Duncan et al.). Scientists also encounter the "black box" issue when using AI in GWAS.

The "black box" issue refers to scientists' difficulty understanding how AI develops its decisions or predictions. While AI is efficient in analyzing complex relationships in data, it does not provide an explanation of which specific genetic variants were significant in coming to its conclusion. This is particularly true with deep learning. Because of the complex layers of neural networks utilized within deep learning, it becomes increasingly difficult for scientists to examine the reasons behind AI-generated predictions or decisions. Within clinical settings, clinicians and

regulators must have a clear rationale behind their predictions or diagnoses, as they cannot simply tell the patients and the FDA to trust the AI. This also raises legal and ethical concerns, as should the model make an incorrect or harmful prediction or decision, and scientists cannot explain why, it becomes more difficult to assign accountability for the mistake, especially if the prediction influences access to treatment or participation in clinical trials. To address these interpretability challenges, researchers are developing explainable AI (XAI) techniques. Methods such as SHAP (SHapley Additive explanations) values quantify each genetic feature's contribution to predictions, while in-silico mutagenesis experiments systematically alter variants to observe their individual effects (Lundberg and Lee; Zhou et al.). Attention mechanisms in neural networks can also highlight which genomic regions the model prioritizes, providing biological insight into model behavior (Lundberg and Lee).

AI is man-made and therefore susceptible to error. Despite AI models' exceptional predictive performance in AD research, they are still in need of rigorous validation in clinical settings. A particular risk is overfitting, where models learn not only meaningful patterns but also random noise from training data, causing them to perform well during development but fail when applied to new datasets. This can produce false positives or missed risk variants in clinical settings, leading to inappropriate treatment decisions. The risk is especially acute when genetic variants are in linkage disequilibrium (LD), where correlations between nearby SNPs can be mistaken for causal relationships (Vilhjálmsón et al.). Without external replication, functional testing, and regulatory approval, these AI models will not be implemented within real-world settings and risk remaining academic exercises rather than actual clinical solutions. Replication on diverse datasets, biological testing of SNP effects, and LD-aware cross-validation techniques are all ways scientists have made efforts to ensure that models do not provide overgeneralized conclusions by overfitting training data, failing in underrepresented populations, or overlooking rare variants (Vilhjálmsón et al.). External test cohorts are essential for validating model accuracy beyond training environments. Additionally, specific training models require data not present within clinical settings, which poses another constraint to AI integration. Ultimately, without this rigorous testing, AI's potential in AD remains largely theoretical.

Data privacy also presents challenges for AI integration in genomics. Genetic data is highly sensitive, and regulatory frameworks like GDPR impose strict limitations on how it can be shared across institutions. This limits the size and diversity of training datasets. Federated learning offers a mitigation strategy by allowing models to be trained across multiple institutions without centralizing sensitive data—the model travels to where data resides and only shares model updates, not raw genetic information, preserving patient privacy while enabling collaborative model development (Rieke et al.; Calvino et al.).

3. Integration of AI Technology in the Context of AD

3.1 AI-Enhanced GWAS for AD

GWAS on its own has shown to be effective within the field of genetics; however,

alongside AI, GWAS could prove to be even more revolutionary. Traditional GWAS methods have shown scientists risk loci for disease; however, with AI, scientists will gain further insight into how each SNP impacts gene and protein functions. Deep learning models, AlphaFold, and ML classifiers are all examples of AI technologies that can be integrated into GWAS and used to improve GWAS performance and enhance the quality of the data scientists collect from GWAS findings. ML classifiers can be trained on known pathogenic variants to identify likely deleterious mutations from millions of SNPs. Deep learning models such as DeepSEA or ExPecto can then aid in the prediction of the effects of non-coding variants on gene expression. Meanwhile, AlphaFold can be used to model the effects that missense variants have on protein function. The ability to predict the effects of missense variants has been useful in AD research, as it has allowed scientists to be able to do so for various genetic variants involved in AD, such as APOE ϵ 4, TREM2, and ABCA7. These predictions, incorporated with multi-omics data, aid in the construction of a comprehensive view of how genetic risk translates into biological dysfunction. AI has great potential within AD, as it can speed up the identification of possible drug targets and reduce the time and cost of GWAS by narrowing down the list of likely genetic variants to a concise list of biologically significant candidates.

3.2 Case Studies: AI in Predicting Risk Variants and Holistic Models

Artificial intelligence has made a significant impact on Alzheimer's disease research and genetics as a whole. The following case studies showcase the effectiveness of AI usage within real-world genetics and AD research.

In a 2021 study conducted by Monk et al., researchers developed a deep learning model called netSNP. This DL model was trained on exome sequencing data collected from nearly 10,000 individuals and identified hundreds of SNPs indicating protective or risk associations of Alzheimer's disease. The number of SNPs was particularly notable because of their correlation with age of onset and neuropathological features. This case study showcases AI's ability to replicate known findings and discover new, significant genetic variants that traditional GWAS methods had missed (Monk et al.).

In another study, Jo et al. created a deep learning method called SWAT-CNN. This method works by dividing the genome into fragments and applying Convolutional Neural Networks to detect phenotype-associated fragments. From there, it then uses Phenotype Influence Scores and a sliding window association test to distinguish relevant SNPs. This model used data from 981 individuals (650 cognitively normal, 331 AD), collected by the Alzheimer's Disease Neuroimaging Initiative (ADNI), and achieved an area under the curve of 0.82 for disease classification (Jo et al.).

Mirabnահrazam et al. conducted another study in which they combined MRI imaging features with genetic data to create a dementia score that predicts the long-term progression of Alzheimer's disease in patients within the ADNI cohort. They compared the predictions from the MRI data, genetic data, and multimodal data. Cognitively normal subjects exhibited a genetic score predicted progression with an accuracy of 0.857, while MRI alone was much lower

(0.143). In cases of mild cognitive impairment, MRI performed better (0.614) than genetics (0.356); however, performance improved in several subgroups when both data types were used. This demonstrates how integrating genetic risk alongside imaging can ultimately improve predictive performance in early stages (Mirabnahrzham et al.).

Collectively, these various case studies showcase the significant strides made with AI's implementation within Alzheimer's disease genomics. AI has allowed researchers to go beyond simple associations and observe complex patterns that can aid in early diagnostic and therapeutic development.

3.3 Potential Impact

One of the most promising impacts of integrating AI with GWAS and other omics data is its ability to improve drug target identification within Alzheimer's disease. By integrating protein interaction, functional omics, and drug-target databases, AI has the potential to identify which genes are viable drug targets, in contrast to traditional GWAS methods, which only identify loci associated with disease risk.

One study illustrates this by combining GWAS findings with multi-omics datasets and the human protein-protein interactome to create a network-based AI framework. This method identified 103 Alzheimer's risk genes and, through retrospective patient cohort validation, identified that drugs such as pioglitazone, febuxostat, and atenolol were significantly associated with reduced AD risk (Cheng et al.). Whereas another study identified the genes INPP5D and PTK2B to be prime target candidates for drug-targeting and identified tamoxifen and bosutinib among repositionable drugs that could modulate Alzheimer's disease-related pathways (Fang et al.). These studies show how AI has been improving the efficiency of genomic drug target identification and allowing for progression within AD research. However, these predictions still need experimental validation, as not all targets are safe or effective in humans. Scientists also encounter the issue of bias here based on the data they train on.

Moreover, AI can be used to enable personalized medical approaches in AD. AI has the potential to be used to tailor treatments, prevention, and diagnosis based on a patient's clinical and genetic profiles. This would involve observing a patient's biomarker levels, polygenic risk scores, imaging data, and lifestyle/environmental data to stratify patients into various subgroups by risk or subtypes and selecting or optimizing their treatment accordingly. Personalized medicine approaches could allow for early intervention and treatment for AD patients, as well as enable clinicians to select more effective treatments for patients based on the data provided to the AI model. There are challenges that are faced with implementing AI in personal medical approaches, such as preserving patient privacy and, again, the potential for data bias.

4. Summary of Current Findings

Genome-wide association studies (GWAS) have aided in the identification of many key SNPs associated with Alzheimer's disease (AD), such as APOE ϵ 4, TREM2, PICALM, and BIN1. In addition, GWAS has aided in the discovery of coding and noncoding variants that

influence a variety of biological processes such as immune response, synaptic function, and lipid metabolism. While GWAS has been essential in locating these genetic variants, scientists struggle to determine their biological significance using traditional methods alone.

Recent integration of artificial intelligence within GWAS has shown significant improvement in its efficiency. AI models such as netSNP and SWAT-CNN have successfully identified known and novel SNP associations by analyzing complex patterns in high-dimensional genomic data. Furthermore, tools such as DeepSEA and AlphaFold have allowed researchers to predict how the missense variants will affect gene expression and protein function, which allows them to more closely observe disease mechanisms.

Many of these AI-driven findings have been rigorously validated using clinical, genetic, and neuroimaging datasets, demonstrating their potential within early diagnostics and risk prediction. However, many challenges remain, such as population bias in training datasets and the integration of findings into therapeutic development. These research gaps emphasize the need for continued research and validation on AI-based methods before implementation within clinical settings.

5. Discussion

5.1 Advantages of AI Integration in GWAS for Alzheimer's Disease

Artificial intelligence has opened up new possibilities in genomics research by enabling the integration of diverse biological data. While GWAS only focuses on the statistical associations of SNPs and diseases, the combination of AI and multi-omic data, such as genomics, proteomics, and neuroimaging, allows researchers to gain further insight from the complex genetic data, such as the interactions, patterns, and effects of genetic variants. This holistic review allows for further understanding of how genetic risk translates into biological dysfunction in AD.

AI can help move GWAS beyond association into biological interpretation by aiding in not only the recognition of associated SNPs, but the analysis of the influence specific variants may have on molecular function. In Alzheimer's disease research, this means identifying how variants may have disrupted pathways in areas such as immune response, lipid metabolism, or neuronal signaling. This allows researchers to extract deeper biological meaning from biological risk variants and is a critical step towards identifying viable drug targets or therapeutic interventions.

Additionally, AI can aid in the development of personalized and precision medicine based on patients' unique biological profiles. By integrating clinical, lifestyle, and multi-omic data, AI can aid in the stratification of patients into subtypes based on their biological profiles and curate unique treatment plans or ensure early interventions tailored to the patient's specific risk factors, disease progression, or therapeutic response. For example, AI can detect subtle molecular patterns that might predict how a patient will respond to a particular drug, reducing trial-and-error in treatment selection and improving outcomes.

5.2 Critical Considerations for AI in Alzheimer's Genomics

While AI has made great strides within genomics, it is certainly not without limitations. As AI tools are being slowly incorporated into GWAS, it is important to address data privacy, model transparency, and reliability concerns. Without proper oversight, AI could raise ethical challenges in genomics, potentially limiting its clinical impact. It is imperative that these concerns are addressed in order to ensure that AI integration is not only innovative but also trustworthy, equitable, and applicable.

AI models in genomics rely heavily on very sensitive personal data, including genetic, clinical, and behavioral information. Should this data not be handled with care or proper consent processes, the patient's private information could be compromised. Furthermore, GWAS datasets upon which the AI models are trained are heavily skewed towards individuals of European descent. This raises ethical concerns around inclusion and fairness, as AI models may perform poorly in underrepresented populations, potentially reinforcing healthcare disparities and limiting the benefits of AD detection for all groups.

While AI models may be proficient in providing predictions, they do not provide clear explanations for them, leading to the findings being dismissed regardless of accuracy. Clinicians and researchers cannot validate predictions without a clear understanding of the processes behind those conclusions. If a model does not provide clinicians and researchers with transparency, the conclusions provided will be deemed uninterpretable, and without interpretability, AI models will remain limited in both scientific and medical communities.

Another concern is overfitting. This occurs when the AI performs well using the training data but falls short when tasked with generalizing new or unseen data. This occurs when AI learns the significant patterns within a dataset, but also the random noise within the dataset. As a result, a model may appear extremely accurate while in development, but when tested on new data, it may produce unreliable predictions as it is unable to replicate the level of accuracy presented with the original data. In a clinical setting, this can lead to false positives or missed risk variants, ultimately reducing the model's efficiency within real-world diagnosis or treatment planning. To avoid this, AI models must be rigorously validated using independent datasets and tested for generalizability.

Future Research Opportunities

Artificial intelligence shows great potential within genomics; however, various areas need further research and development in order to optimize its efficiency within the field. Researchers must begin diversifying genetic data sets to include ancestral populations in order to improve AI's generalization capacity. Researchers must prioritize recruiting participants from underrepresented ethnic groups and train models with these datasets in order to aid in the identification of population-specific risk variants. This can be accomplished through collaboration amongst international consortia, similar to the International Genomics of Alzheimer's Project (IGAP), but with the specific goal of diversity recruitment. Researchers must also focus on integrating longitudinal multi-omics data, drawing from a combination of

genomics, transcriptomics, metabolomics, and proteomics with clinical measurements over time. In order to do this, scientists would need to develop standard data collection protocols across various research centers as well as develop AI models that are capable of analyzing temporal patterns in biological changes that precede the cognitive decline that comes with Alzheimer's disease.

Another promising direction involves the improvement of model interpretability through explainable AI (XAI) techniques. This could be done by incorporating biological pathways into neural network structures or using attention mechanisms to identify which genetic variants could have the strongest impact on predictions. This could also involve creating hybrid models of deep learning and knowledge of graphs of protein structures and regulatory networks. Doing this will eliminate the “black box” issue previously mentioned, as it would allow scientists to not only see AI predictions but also how the AI model was able to come to those conclusions. However, the use of prospective validation studies is also essential—researchers should design clinical trials that utilize AI-derived polygenic risk scores to stratify patients for early intervention, tracking outcomes over 5-10 years. This would allow researchers to assess whether or not AI-guided precision medicine approaches will actually improve patient outcomes. These validation studies would require collaboration between computational researchers, clinicians, and regulatory bodies to establish appropriate endpoints and safety monitoring protocols.

Conclusion

Alzheimer's disease is a complex neurodegenerative disorder driven by a mix of genetic factors and biological changes. Traditional genome-wide association studies (GWAS) have helped pinpoint genetic variants linked to the disease, but they often struggle to explain how these variants actually affect biology. The use of artificial intelligence has transformed this process, making it faster and more precise to identify and interpret relevant genetic markers. AI's ability to handle large, complex datasets uncovers patterns that deepen our understanding of how genetics contributes to Alzheimer's. As these technologies advance, they could improve early detection, risk assessment, and help tailor treatments to individuals. While challenges remain, AI marks an important step toward better understanding and eventually finding effective therapies for Alzheimer's disease.

Works Cited

- Bird, Thomas D. "Genetic Aspects of Alzheimer's Disease." *Genetics in Medicine*, vol. 10, no. 4, Apr. 2008, pp. 231-239.
- Buniello, Annalisa, et al. "The NHGRI-EBI GWAS Catalog of Published Genome-Wide Association Studies, Targeted Arrays and Summary Statistics 2019." *Nucleic Acids Research*, vol. 47, no. D1, 8 Jan. 2019, pp. D1005-D1012.
- Calvino, Giulia, et al. "Federated Learning: Breaking Down Barriers in Global Genomic Research." *Genes*, vol. 15, no. 12, 22 Dec. 2024, <https://www.mdpi.com/2073-4425/15/12/1650>. Accessed 23 Oct. 2025.
- Cheng, Feixiong, et al. "Artificial Intelligence Framework Identifies Candidate Targets for Drug Repurposing in Alzheimer's Disease." *Alzheimer's Research & Therapy*, vol. 14, no. 7, 10 Jan. 2022, <https://alzres.biomedcentral.com/articles/10.1186/s13195-021-00951-z>. Accessed 10 Oct. 2025.
- Duncan, Lauren, et al. "Analysis of Polygenic Risk Score Usage and Performance in Diverse Human Populations." *Nature Communications*, vol. 10, no. 3328, 25 July 2019, <https://www.nature.com/articles/s41467-019-11112-0>. Accessed 23 Oct. 2025.
- "Early-Onset Alzheimer's Disease." *Johns Hopkins Medicine*, 15 Apr. 2024, <https://www.hopkinsmedicine.org/health/conditions-and-diseases/alzheimers-disease/early-onset-alzheimer-disease>. Accessed 10 Oct. 2025.
- Fang, Jiansong, et al. "Artificial Intelligence-Based Computational Framework for Drug-Target Prioritization and Inference of Novel Repositionable Drugs for Alzheimer's Disease." *Alzheimer's Research & Therapy*, vol. 13, no. 92, 3 May 2021, <https://alzres.biomedcentral.com/articles/10.1186/s13195-021-00826-3>. Accessed 10 Oct. 2025.
- Jo, Taeho, et al. "Deep Learning-Based Identification of Genetic Variants: Application to Alzheimer's Disease Classification." *medRxiv*, 19 July 2021, <https://www.medrxiv.org/content/10.1101/2021.07.19.21260789v2.full-text>. Accessed 10 Oct. 2025.
- Levy, Daniel. "Three Billion Base Pairs vs. One Powerful Computer." *NIH Intramural Research Program*, Nov. 2018, <https://irp.nih.gov/blog/post/2018/11/three-billion-base-pairs-vs-one-powerful-computer>. Accessed 10 Oct. 2025.
- Lundberg, Scott M., and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4765-4774, https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf. Accessed 23 Oct. 2025.
- Mirabnahrzham, Ghazal, et al. "Machine Learning Based Multimodal Neuroimaging Genomics Dementia Score for Predicting Future Conversion to Alzheimer's Disease." *Journal of Alzheimer's Disease*, 2022, <https://journals.sagepub.com/doi/abs/10.3233/JAD-220021>. Accessed 10 Oct. 2025.

- Monk, Bradley, et al. "A Machine Learning Method to Identify Genetic Variants Potentially Associated With Alzheimer's Disease." *Frontiers in Genetics*, vol. 12, 13 June 2021, <https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2021.647436/full>. Accessed 10 Oct. 2025.
- Rieke, Nicola, et al. "The Future of Digital Health with Federated Learning." *npj Digital Medicine*, vol. 3, no. 119, 14 Sept. 2020, <https://www.nature.com/articles/s41746-020-00323-1>. Accessed 23 Oct. 2025.
- Sollis, Elliot, et al. "The NHGRI-EBI GWAS Catalog: Knowledgebase and Deposition Resource." *Nucleic Acids Research*, vol. 51, no. D1, 6 Jan. 2023, pp. D977-D985.
- Uffelmann, Emil, et al. "Genome-Wide Association Studies." *Nature Reviews Methods Primers*, vol. 1, 26 Aug. 2021, <https://www.nature.com/articles/s43586-021-00056-9>.
- Vilhjálmsón, Bjarni J., et al. "Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores." *The American Journal of Human Genetics*, vol. 97, no. 4, 1 Oct. 2015, pp. 576-592, [https://www.cell.com/ajhg/fulltext/S0002-9297\(15\)00365-1](https://www.cell.com/ajhg/fulltext/S0002-9297(15)00365-1). Accessed 23 Oct. 2025.
- "What Are the Signs of Alzheimer's Disease?" National Institute on Aging, <https://www.nia.nih.gov/health/alzheimers-symptoms-and-diagnosis/what-are-signs-alzheimers-disease>. Accessed 10 Oct. 2025.
- "What Happens to the Brain in Alzheimer's Disease?" National Institute on Aging, 19 Jan. 2024, <https://www.nia.nih.gov/health/alzheimers-causes-and-risk-factors/what-happens-brain-alzheimers-disease>. Accessed 5 Sept. 2025.
- "What is Alzheimer's Disease? Symptoms & Causes | alz.org." Alzheimer's Association, <https://www.alz.org/alzheimers-dementia/what-is-alzheimers>. Accessed 6 Sept. 2025.
- Wightman, Douglas P., et al. "A Genome-Wide Association Study with 1,126,563 Individuals Identifies New Risk Loci for Alzheimer's Disease." *Nature Genetics*, vol. 53, no. 9, 2021, pp. 1276-1282.
- Yamazaki, Yu, et al. "Apolipoprotein E and Alzheimer's Disease: Pathobiology and Targeting Strategies." *Nature Reviews Neurology*, vol. 15, no. 9, Sept. 2019, pp. 501-518.
- Zhou, Jian, et al. "Deep Learning Sequence-Based Ab Initio Prediction of Variant Effects on Expression and Disease Risk." *Nature Genetics*, vol. 50, no. 8, 16 July 2018, pp. 1171-1179, <https://www.nature.com/articles/s41588-018-0160-6>. Accessed 23 Oct. 2025.

High School Educator Perceptions on Implementing Italian Education Practices Into the American Traditional Education System By Paige Bruskrud

Abstract

The purpose of this study was to understand high school educator perceptions on implementing Italian education practices into the American traditional education System. This mixed method study employed a pre-test and post-test design with high school educators to determine strengths and weaknesses of the traditional education system. Teachers (N=12) completed an interview with the researcher where they responded to Likert scale survey questions (pre/post-test) along with a series of interview questions that differed pre/post-test. The Likert scale survey responses underwent Pearson Correlation Coefficient tests and pre/post-test histograms were subsequently created. For the qualitative interview responses, the teachers' answers were coded based on theme and sorted into a thematic analysis table to be analyzed further. Results concluded that educators do strongly believe that Italian education could be successfully implemented into American traditional education to address weaknesses regarding structural barriers and autonomy. Implications for this study include how Italian education elements can be incorporated into American traditional education and how valuable instructors' voices are in the journey for educational reform.

Introduction

The American education system, traditionally characterized by standardization, efficiency, and measurable outcomes, faces growing criticism for its inability to individualize student development. Previous state and federal mandates, including the No Child Left Behind Act (2001) and Race to the Top (2009), require that public schools receiving federal funding administer statewide standardized testing consecutively. Furthermore, schools receiving Title I funding through the Elementary and Secondary Education Act (1965) must show yearly progress on academic test scores. Due to these legislative acts, the purpose of education has become narrowed down to “teaching to the test,” which neglects other important educational goals (Corrigan). Additionally, 51% of American adults say public K-12 education is going in the wrong direction due to issues such as funding and resource challenges; the need for educational reform is paramount (Minkin).

The purpose of this study is to address alternative educational approaches that can be integrated into the American education system. The examined approaches include Italian education (IE) models, Reggio Emilia, and Montessori, which focus on student autonomy, inquiry, and hands-on learning. These approaches not only support academic achievement but also foster adaptability, social collaboration, and a deeper engagement with learning. Integrating key elements of IE philosophy into American schooling has the potential to revitalize classroom environments for both teachers and students to prepare future generations for the demands of a rapidly evolving educational landscape (Rathunde).

While this study is adaptable to all levels of education, the concentration will be on the

high school demographic as the researcher is a high school junior with a vested interest in the research. This paper seeks to answer the research question: How can IE practices be implemented into the American high school education system to address weaknesses within traditional education through teacher perspectives?

Literature Review

Background and Pedagogy of IE

Reggio Emilia education emerged after World War II, when parents envisioned schools centered on the image of the child. With political support, Reggio Emilia opened its doors in 1963 as the city's municipal preschool (Kristiyani). Martalock conducted a literature review on Reggio Emilia education, discovering that the most significant component was its project-based approach where "children engage together, along with the teacher, in an on-going process of sharing ideas and building common understandings" (7). The Reggio Emilia approach is centered around the image of the child leading to flexible engaging classroom environments, teachers acting as researchers, and active learning. Additionally, Martalock curated tables to create distinctive side-by-side comparisons to American traditional education (ATE) which aligned with this research paper's educational intervention.

Montessori education, developed by physician Maria Montessori, emphasizes hands-on, sensory learning to build social and independent skills. In 1907, Montessori opened Casa dei Bambini (Children's House), and its success quickly gained global recognition (Kristiyani). To further explore Montessori practices, research conducted by Aljabreen found that Montessori teachers serve as guides, balancing child-led interests with teacher-supported instruction. Children are encouraged to develop independence early on, working at their own pace based on their interests and skills. Teachers intervene minimally during peer conflicts to promote emotional self-regulation and independence. Furthermore, the classroom environment is multi-aged and structured to engage students in real world skills such as cooking, cleaning, or dressing, further encouraging independent work.

Efficacy of IE

The effectiveness of IE, specifically Montessori education, can be demonstrated through a study conducted by Rathunde. The purpose of the study was to provide positive reinforcement for Montessori education by applying it to the broader context of ATE through an Experience Sampling Method (ESM). Results revealed that Montessori students felt more active, excited, and motivated doing academic work compared to the ATE group. Moreover, Montessori students experienced both high motivation and high salience at the same time, whereas ATE students only experienced this 24% of the time. As described by Rathunde, with public schools continuing towards a path of top-down education and standards-based testing, reform efforts have been futile. However, implications from this study include how "the benefits associated with the

Montessori schools, therefore, provide evidence that supports the case for reform,” however as the study was published over 20 years ago this assertion may not be applicable.

Another study that advances the findings of Rathunde’s research is Randolph et al. The purpose of this study was to examine the impact of Montessori education on academic and non-academic outcomes when compared to ATE. Using a meta analysis of 19 academic databases in Montessori related journals, researchers chose articles that directly compared the two forms of education. Results of the study concluded that Montessori education exemplified a positive and meaningful impact on a child's outcomes. The study found that by first grade, Montessori students gained about two extra months of learning, and by sixth grade, the advantage grew to roughly an extra year. However, the strongest benefits were in non-academic areas where Montessori students showed significantly greater self-control, focus, and problem-solving skills compared to ATE students. This highlights a commonality between the studies, as Montessori students are more intrinsically motivated than ATE students. Despite the constant evolution of the world, the efficacy of Montessori education has not changed since Rathunde’s study, as proven by Randolph et al.'s similar conclusions.

IE Implementation

The successful implementation of IE concepts into ATE settings was demonstrated by Steglin who applied Reggio Emilia principles to an elementary school science lesson. Using the Benchmark Content Standards and state-mandated requirements, Steglin provided a curriculum framework for teachers to implement while creating an engaging learning environment. The process began with teacher inquiry into student interests through observation of peer interactions in informal settings or lunch/nutrition time. From there, the teacher identified a common student interest in rainforests to develop benchmark-aligned activities, such as illustrating rainforest life cycles or documenting ecosystem relationships. This lesson plan illustrates how the project-based and student-led interests of IE can be successfully implemented into standardized ATE, which mitigates issues like low engagement and disciplinary challenges. Another implication identified includes how other subjects could be tailored to Reggio Emilia concepts by using this lesson plan as a template.

A study that supports these findings is Fleming et al., who investigated existing Montessori implementation in South Carolina’s public schools. Researchers intensely observed, interviewed, and surveyed the principals and teachers to determine the success of its previous implementation. Findings concluded that despite education standards, programs in South Carolina were able to sustain Montessori practices. Teacher interviews also revealed that, with proper district support, public Montessori schools can continue to thrive. These findings align with Steglin, as both studies found that despite standardization or benchmarks, integration of IE can be successful. This evidence encouraged the researcher of this study to investigate deeper into which methods and practices would achieve successful implementation in ATE schools. The researcher was inspired by the recency and relevancy of Fleming et al.'s research that used semi-structured teacher interviews and Likert-scale survey questions for the methodology.

Identified Gaps

In the literature, the researcher identified three prominent gaps. First, the implementation of IE is frequently discussed in the context of younger age groups. Studies conducted by Martalock, Randolph et al., and Fleming et al. all focus on transitioning to IE for preschool-elementary age groups. According to the National Center for Montessori in the Public Sector, there are only 150 public Montessori high schools in the US, and with no official database, research focused on this group remains limited (NCMPS). A second gap that exists is understanding the most effective way to implement IE within existing ATE schools. Fleming et al. “attempted to gauge the possibility of success for the Montessori model in the public sector” by analyzing the schools already operating under Montessori principles. This leaves a discrepancy in the research, regarding how best to introduce IE into established ATE schools that do not already incorporate Montessori practices. Lastly, Martalock and Randolph et al. studies overlook educators' perspectives, even though teachers are responsible for implementing IE. Kannwischer states that "teachers' beliefs are the best indicators of classroom decision-making and behaviors," highlighting the importance of their input. As such, the study in this paper sought to address these gaps by examining the most effective methods of implementing IE in ATE high schools through the lens of educators.

Methods

A mixed method study that included a quasi-experimental design component was implemented through interviews with teachers from Calabasas High School. Educator perceptions were measured through semi-structured interview questions and Likert-scale survey questions, with an educational intervention between the pre- and post-tests. Specifically, this research method aligns with Fleming et al. regarding teacher perceptions through interviews and surveys.

Population and Sample Selection

The target population for the study were educators from Calabasas High School who teach ATE. Through purposive sampling, 15 educators were selected to participate in the study. This population is different from Fleming et al.'s research that interviewed Montessori teachers specifically, as this researcher interviewed only ATE teachers. The researcher used purposive sampling in order to achieve similarity, which is vital for eliminating any confounding variables in quasi-experimental designs. Similarly, consistent variation in demographics such as grade level taught, gender, and subject taught were considered in the selection process. Additionally, all teachers participated in both the pre- and post-test to remove any potential distortion of results, demonstrating the importance of purposive sampling (Office for Health Improvement and Disparities).

Within the 15-teacher population, a subpopulation of teachers who instruct Humanities subjects (66.7%) versus those who teach STEM subjects (33.3%) was identified. The goal of having this subpopulation (Humanities versus STEM) was to compare continuing and differing

perspectives depending on the subject taught. Within this subpopulation, the disciplines were broken down into VAPA (Visual and Performing Arts), English, History, Math, and Science. Furthermore, to eliminate any confounding variables based on gender, 50% of the population were female teachers and 50% were male.

Implementation

To implement the researcher's design, an email was sent out to all 15 teachers originally chosen to be in the study with details describing expectations for the data collection. Out of these 15 teachers, two respondents declined participation; therefore, the researcher sent out emails to other teachers who taught the same subject. When follow-up was necessary, the researcher talked in person to each instructor individually. However, despite the researcher's best efforts to contact and schedule a meeting with the teachers, due to time constraints, three teachers could not be reached, resulting in a sample size of 12 teachers.

For three months, 30-60 minute in-person interviews occurred. The researcher's method process was similar to Fleming et al., as both researchers collectively utilized a combination of interview and survey questions. These choices were made to obtain both quantitative and qualitative data that could be compared to post-test data. Specifically, the Likert scale questions were important as they provided statistical and generalizable evidence about changing/continuing perceptions of teachers. Likert scale questions such as "Are able to meet the needs of students challenged academically?" and "Are able to meet the needs of the students challenged behaviorally?" from Fleming et al. helped form the researcher's questions. Prior to the pre-test, teachers were required to sign a letter of consent (Appendix A) before moving onto ten Likert scale questions (Table 1). The survey questions gauged participants' views on ATE practices, followed by five semi-structured interview questions (Table 2) that probed further to discern the participants' viewpoint on the topic of ATE before the intervention was provided.

The intervention entailed three infographics explained by the researcher detailing the benefits of IE with a side-by-side comparison to ATE. The researcher described the infographics to each of the participants for about five to ten minutes. This intervention was designed through the app Canva to curate the infographics (Table 4), which were based upon and supported by pre-existing literature such as Martalock. Following the intervention, during the post-test, interviewees were instructed to complete the same Likert scale survey questions from the pre-test to determine if the infographics changed perceptions on ATE. Another set of five semi-structured interview questions (Table 3) were asked only about IE and possible methods of implementation in ATE to answer the research question, marking the conclusion of the study (Leedy and Ormrod 178-210).

To ensure replicability various tables are included in the appendix for the instruments. In Table 1 (Appendix B) the ten Likert Scale Survey Questions (pre-/post-test) are included with three columns detailing each question. In column one is the question on the Google Form, column two depicts the measurement scale used to discern the statistical data, and the third column cites the source the question is derived from.

Following Table 1, is Table 2 (Appendix B) which presents the pre-test semi-structured interview questions asked to the participants by the researchers. This table is organized into three columns: the first states the question posed, the second identifies the ATE topic each question addresses, and the third cites the source the question is based upon. Table 3 (Appendix B) follows the same structure however the interview questions used in this table are specifically for the post-test and focused on participants perceptions on IE and implementation. Table 4 (Appendix B) displays the infographics curated by Canva and supported by Martalock's research. Three infographics were curated with the first one detailing a typical day in Montessori education from start to finish. The next infographic offers a side-by-side comparison of Reggio Emilia education to ATE, touching on topics such as the image of the child and curriculum employed. Lastly, the third infographic describes overall statistical data about IE and a broader discussion into its pedagogy.

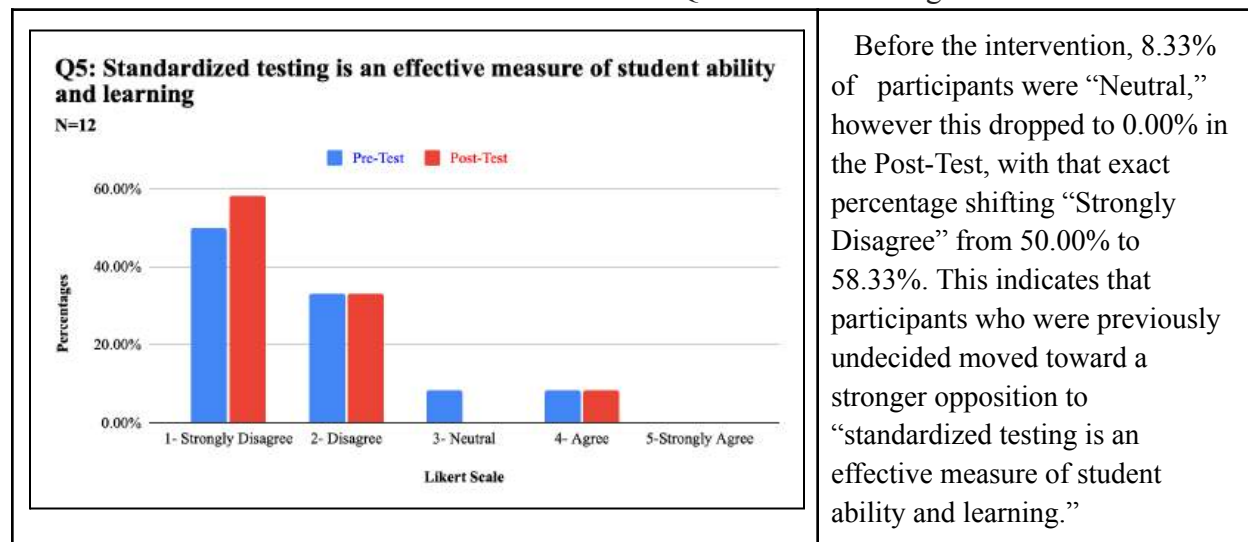
Data Analysis System

Pre- and post-test data assessed the intervention's impact on participants' views of IE and ATE, as well as effective IE implementation methods. Subsequently, the researcher used quantitative analysis for the Likert scale survey responses. The participant survey responses were automatically transferred into Google Sheets, where pre-test/post-test histograms were created, and Pearson Correlation Coefficient tests were conducted using a Social Scientific Calculator and XLMiner Analysis. In addition, the researcher conducted a qualitative analysis by coding themes from transcriptions of the semi-structured interviews in the pre- and post-tests through the transcription software, Cockatoo.

Findings

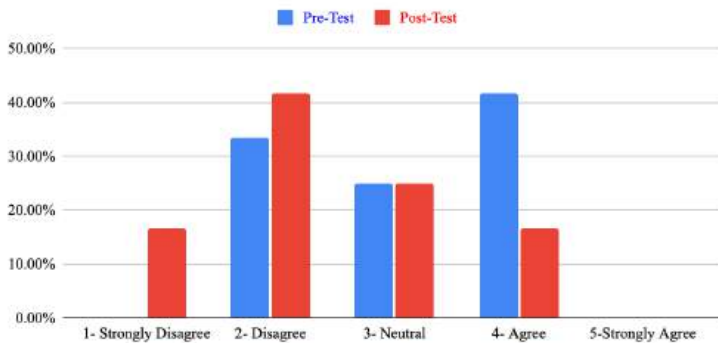
Quantitative Findings

Table 6 Pre-Test/Post-Test Quantitative Findings



Q7: Teachers have enough autonomy in the traditional system to adapt lessons to individual student's needs

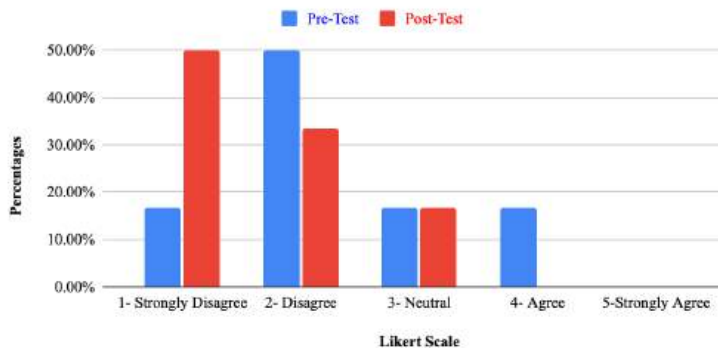
N=12



After the intervention, statistical changes occurred in the “Strongly Disagree,” “Disagree,” and “Agree” sections. In the “Agree” section, 41.67% originally selected this option; however, in the Post-Test, only 16.67% of participants did. The 25% decrease caused increases in both the “Strongly Disagree” and “Disagree” sections during the Post-Test, suggesting that educators began to question their own teacher autonomy after the educational intervention.

Q8: The traditional education system supports teachers motivation and growth in the classroom

N=12



The most notable changes occurred in the “Strongly Disagree,” “Disagree,” and “Agree” sections. In the “Agree” section after the intervention, results changed from 16.67% to 0.00%. Another category where there was a decrease was in the “Disagree” section from 50.00% to 33.33%. These decreases created a large jump in the “Strongly Disagree” sections from 16.67% to 50.00%, demonstrating how after the intervention teachers opinions on the support they receive worsened and scores lowered.

Q10: Classroom environments are aesthetically appealing, friendly, and promotes learning

N=12

Response	Pre-Test (%)	Post-Test (%)
1- Strongly Disagree	0.00%	0.00%
2- Disagree	16.67%	58.33%
3- Neutral	66.67%	33.33%
4- Agree	8.33%	8.33%
5- Strongly Agree	8.33%	0.00%

Initially 8.33% of participants selected “Strongly Agree” in response to the classroom environment question, however this number dropped to 0.00% in the Post-Test. Additionally, the percentage of participants who selected “Disagree” increased from 16.67% to 58.33% between the Pre-Test and Post-Test. These shifts suggest that the intervention may have influenced participants' understanding, leading them to recognize how alternative education methods often provide more appealing learning environments compared to ATE.

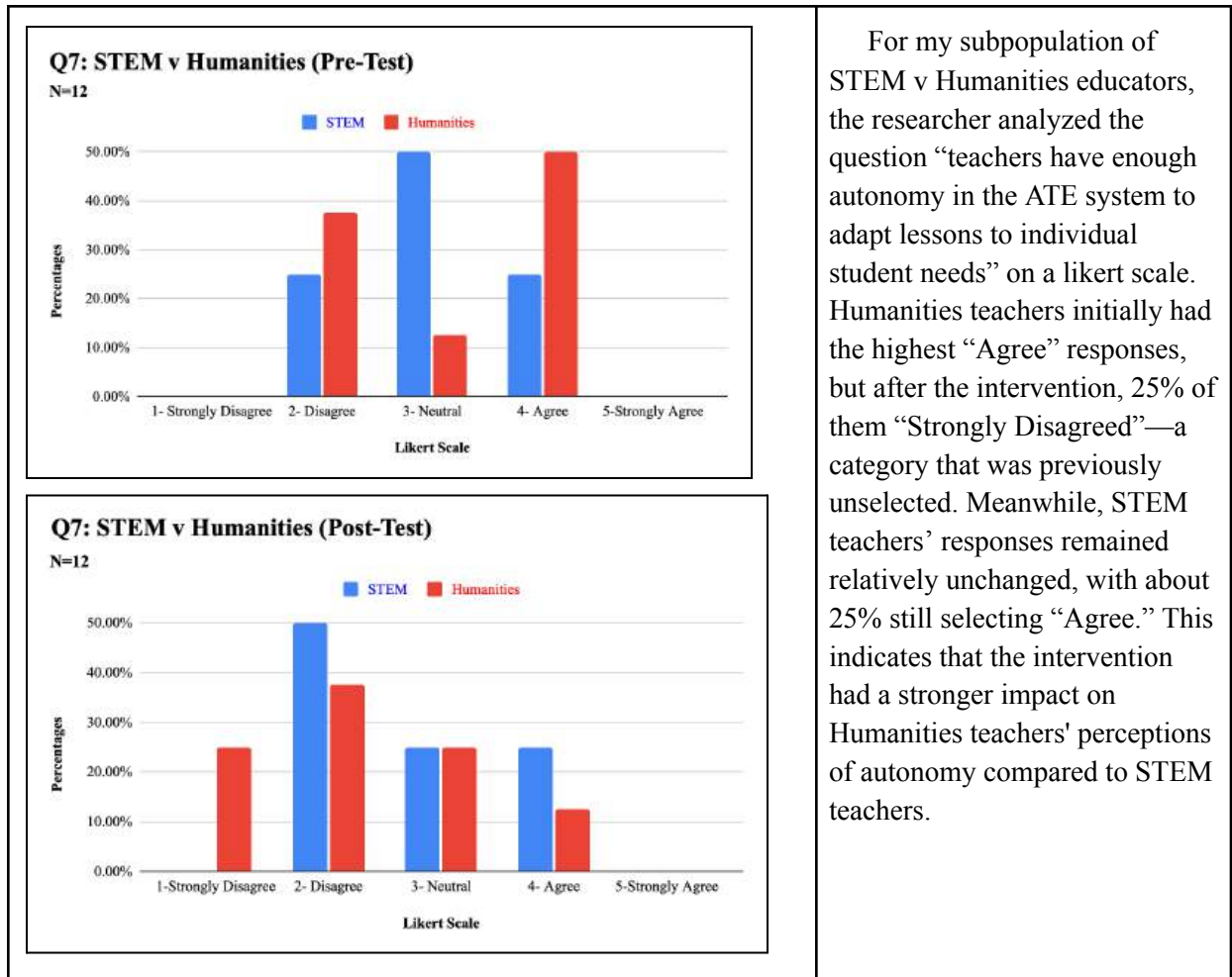
Q11: Traditional learning involves active learning

N=12

Response	Pre-Test (%)	Post-Test (%)
1- Strongly Disagree	0.00%	25.00%
2- Disagree	41.67%	50.00%
3- Neutral	25.00%	16.67%
4- Agree	33.33%	8.33%
5- Strongly Agree	0.00%	0.00%

A majority of the categories changed, with selection of “Neutral” and “Agree” responses decreasing by at least 9% and the “Strongly Disagree” section jumped from 0% to 25.00% after the intervention. The findings suggest that after learning about IE’s hands-on, exploratory methods, participants became more doubtful about ATE.

Table 7. Subpopulation Quantitative Findings (Pre-Test/Post-Test)



Correlational Findings

Table 8. Pearson Correlation Coefficient

Question 3 and Question 4	Values	Significance
r	0.9065420561	Strong Positive Correlation
p	0.000048	*indicates results are significant at <0.05

A Pearson correlation analysis for the Likert scale survey questions illustrated a strong positive relationship ($r = 0.910$, $p < 0.05$) between responses to question 3 (“The ATE structure fosters a love for learning among students) and question 4 (“The ATE system prioritizes student mental health and well being”). As the majority of participants expressed criticism toward the ATE system—such as 83.33% “Disagreeing” or “Strongly Disagreeing” that it fosters a love for learning—the correlation indicates that those who rated one aspect of the system negatively were very likely to rate the other negatively as well. This suggests a consistent perception among participants that the ATE system does not support developmental needs and inspire a love for learning.

Table 9. Pearson Correlation Coefficient

Question 4 and Question 5	Values	Significance
r	0.8161496523	Strong Positive Correlation
p	0.001206	*indicates results are significant at <0.05

For question 4 (“The ATE system prioritizes student mental health and well-being”) and question 5 (“Standardized testing is an effective measure of student ability and learning”) a Pearson Correlation Coefficient test was run. Results indicated that there was a Strong Positive Correlation ($r=.816$) and a statistically significant p-value ($p < 0.05$) between the two questions. With about 50% of participants choosing a “Strongly Disagree” or “Disagree” response to question 4, and 91% for question 5, suggests that those who selected either category for one question were likely to select the same category for the other question. Thus, educators consistently expressed concerns about the ATE system’s failure to prioritize student mental health and the ineffectiveness of standardized testing.

Transcript Thematic Analysis

Using the 12 interviews conducted, the researcher was able to identify the eight most prominent themes. These themes and definitions have been coded into Table 10, below.

Table 10. Definition of Themes

Theme	Definition
Structural Barriers in ATE	Systemic obstacles in ATE high schools including large class sizes, micromanagement from districts, inequitable resource distribution, and funding issues which impact student learning and teacher effectiveness.
Standardized Testing & Curriculum Constraints	Experiences with test and curriculum based systems reveal negative impacts such as students stress, loss of creativity, a focus on rote memorization over critical thinking, and constraints on teacher flexibility.
Teacher and Student Autonomy	Restrictions on teacher and student autonomy in regards to the limited flexibility students have in learning pace, caused by the rigid “one size fits all” approach to ATE.
Alternative & Experiential Learning Approaches	Non-traditional teaching methods such as outdoor education, project-based learning, hands-on learning, and exploratory learning integration.
Classroom/Scheduling Innovations	Methods of implementing IE through changes to classroom environments, structure, and staggered schedules.

Theme	Definition
The Role of Technology in Education	Overdependence on technology in ATE causes teacher ineffectiveness, lack of engagement, and student disciplinary issues.
The Status of Visual & Performing Arts (VAPA)	Examines the undervaluing of Visual & Performing Arts (VAPA) in schools which includes limited funding and reduced arts education prompting the need for prioritization of the subject.
Reform & Teacher-Led Change	Methods of reform in order to implement IE along with teacher led initiatives. Includes openness/change in perception to IE as well as pilot programs and educating other teachers.

The results and summary from the teacher interviews were sorted into Table 5 (Appendix C) with the five emergent themes for each interview being identified. As stated before, 12 high school teachers from Calabasas High School were interviewed using a mixed-method quasi-experimental design to determine potential integration of IE, and shifts in viewpoints. In order to maintain confidentiality, participant names were not listed; hence, they were given a participant ID number following the order in which they were interviewed. The second column of the table contains the specific themes identified and defined in Table 10 within each interview. The third column is a summary of the main details that correlate with the theme within the row. From the interview transcripts, the researcher was able to select the most significant experiences and points that the interviewees mentioned. This enabled the researcher to recognize the multitude of subthemes that exist within the one generalized theme to define the themes within Table 10. For example, in Interview 5, the participant discussed how “the biggest hindrance to [her] is class size,” as “it’s not even just about in the moment interactions; it’s the workload outside of teaching” (Interview 5 17:59). Her mention of class sizes corresponds with the theme “Structural Barriers in ATE” as it addresses how the structure of large classes in ATE is a major barrier. Therefore, these subtopics demonstrate the varying experiences and issues that catalyze participants’ views within each theme. Moreover, the fourth column contains the timestamps of the interview to document when themes began to emerge during the interview.

Results

The interviews conducted revealed educator perceptions of prevalent issues, experiences, and potential implementation methods within ATE high schools. Table 10 identifies themes and definitions from the interviews, while Table 5 corresponds these themes with interview summaries (Appendix D). This section will discuss the findings from the thematic analysis in order to answer the research question and address gaps.

Recurring Theme 1: Structural Barriers in ATE

As defined in Table 10, Theme 1 includes systemic issues regarding large class sizes, micromanagement from districts, inequitable resource distribution, and funding issues. This

theme trended in 83.33% of interviews with educators stating the most significant issue is large classroom sizes. Participants in Humanities or STEM also mentioned how large class sizes were a significant barrier, demonstrating a school-wide challenge rather than a subject-specific one. For instance, participant 10 expressed how

“There are certain things to consider that we don’t really have control over, like class sizes, . . . where it’s really difficult to do some of these labs or do more of these group activities or give more individual-based instruction because there’s just too many kids” (Interview 10 6:46-7:02).

This information is beneficial as it directly addresses the portion of the research question focused on weaknesses in ATE from a teacher's perspective by identifying how large class sizes are the most significant weakness. Ultimately, high-volume class sizes make it difficult for teachers to accommodate and provide individualized support to students. Moreover, for subjects like science as discussed by the interviewee, larger class sizes mean that funding for labs and individual student equipment is difficult, exacerbating inequitable funding issues.

Recurring Theme 2: Teacher and Student Autonomy

As stated in Table 10, Teacher and Student Autonomy can be defined as restrictions on their autonomy in regards to the limited flexibility caused by a rigid “one size fits all” approach in ATE. Additionally, 83.33% of participants expressed how autonomy in the classroom is limited due to structural barriers previously discussed. Often, teachers are forced to default to ATE teaching methods such as lectures and homework assignments which disengage students, resulting in disciplinary issues and lowered academic outcomes (Interview 6 Technology Issues). Lowered academic outcomes stem from varying student learning paces, making it impossible for all students to keep up with academic rigor. In essence, participants 5, 8, and 11 described ATE as “black and white” and extremely “one size fits all,” with Participant 2 saying, “one size fits all ends up being one size fits none” in reference to ATE (Interview 2 3:05-3:10).

Similar to the theme of Structural Barriers in ATE, both STEM and Humanities participants discussed issues with ATE high schools that extend beyond individual disciplines and into the entire system. Another similarity is that teachers described how they are powerless in mitigating large class sizes and the one-size-fits-all mold, demonstrating systemically wide obstacles.

Recurring Theme 3: Alternative and Experiential Learning Approaches

75% of participants mentioned the theme of non-traditional teaching methods. Before with the previous themes, educators described implementation of IE practices improbable due to those barriers. However, after the researcher’s intervention, multiple teachers discussed how they could implement Italian practices such as outdoor education, project-based learning, and hands-on learning. One example of this is from participant 5, a STEM teacher, who said:

“You give me a good idea now, because you know how in a 45, 45, 90 triangle, like we go to the baseball field and show, and you're talking a little about baseball, why more runners will try to steal second base and third base because it's a longer throw, and you'll get to actually go out there and measure” (Interview 4 24:33-24:49).”

This moment was following a discussion where participant 2 mentioned how he wants to implement the outdoor environment aspect of IE following the researchers' intervention. This could imply that the intervention did have a positive effect on this teacher's mindset towards implementation, as he was inspired by the Italian system's real-world application practices.

Additionally, the STEM versus Humanities educators revealed distinctions in preferred learning approaches. STEM educators demonstrated interest in a real-world application approach, with participant 7 saying she would want to “take a field trip to the beach to look at the waves and look at the sinusoidal motion” (Interview 7 19:01-19:27). The interest in real world application however was not favored by Humanities subject teachers who cited experiential, project-based, and hands-on learning, implying that within the umbrella of Alternative and Experiential Learning Approaches there is a teaching method suitable for every subject that could be successfully adapted.

As these aforementioned teaching methods are rooted in Italian pedagogy this could suggest that American high schools could implement components of Italian education through varying teaching methods. Another significant finding included how all Humanities and STEM subject teachers mentioned how having an outdoor education or environment component is something they want to incorporate due to positive benefits to student mental health.

Recurring Theme 4: Classroom/Scheduling Innovations

As defined in Table 10, Classroom/Scheduling Innovations are techniques to integrate Italian education through transformations to classroom environments, structure, and staggered schedules. In 75% of the interviews, the researcher discovered participants' insights into this theme. Within this, participants mentioned how they believe scheduling innovations such as staggered schedules or adjustments to the length of school days are the most effective paths to implementation. A Humanities teacher spoke to this, stating:

“I also don't think that the school day has to be as long. Like I think you could have more potent learning in a short amount of time than just like ‘oh we need to fulfill this many hours for the school year.’ Like I think you can get better quality learning done in a shorter amount of time” (Interview 8 31:37-32:03).

Essentially, the teacher is explaining how school days should be shorter, as it allows for a better quality of learning. This sentiment was also shared by participants 7 and 11, who believed that a reduction in class periods would lead to a positive impact on student mental health and

engagement by alleviating stress. Moreover, all these participants agreed, hypothetically, that a staggered schedule with a morning/afternoon block would be the most effective way to integrate IE. These participants believed staggered schedules would allow for smaller class sizes which they viewed as the ultimate hindrance in ATE. Moreover, both Humanities and STEM teachers shared alike views in this scenario; however, the majority of the STEM teachers favored a schedule innovation rather than a classroom one.

Recurring Theme 5: Reform/Teacher-Led Change

Lastly, 66.67% of participants referenced this theme consistently, regarding methods of reform, in order to implement IE along with teacher-led initiatives. Interviews 3 and 5 discuss pilot programs beginning at the Kindergarten level to developmentally build towards the high school level. One of the Humanities teacher stated:

“To build the framework first and for that you have to start in kindergarten. Yeah because if you introduce something to fifth graders if they've never done it for the last five years why would they know what they're doing?” (Interview 3 16:23-16:50)

Succinctly put, this participant was referencing how the success of pilot programs must first be integrated at the Kindergarten level until it reaches high school. Similarly, a STEM teacher believed that a pilot program must be administered at the Kindergarten level, saying that it would have to begin at one school to serve as a testing ground to be adapted elsewhere (Interview 5, 27:49-28:00). As these are both STEM and Humanities participants, it exemplifies a common viewpoint. Additionally both participants have been teaching for 1-3 yrs, suggesting that teachers who are new to the field of education can offer valuable insight and fresh perspective into educational reformation. The other six participants interviewed expressed implementation methods relating to teacher-led initiatives. They suggested teacher collaboration or educating teachers during faculty/department meetings are effective ways to integrate IE.

Addressing the Gaps

These findings address the gap in the literature, as there is a lack of prior research such as Fleming et al.'s, discussing ways to implement IE into existing ATE schools. Thus, these participants answer the gap and, by extension, the research question as they provide details on the most successful ways to integrate IE into ATE. As described above, a majority of participants believed that the most appropriate integration methods included Classroom/Scheduling Innovations and Reform & Teacher-Led Change.

Discussion

Similar to the researcher's Likert scale survey findings, Fleming et al. discovered teacher surveys indicated the belief that standardized testing poses major challenges to integration, with researchers arguing these tests are inadequate and inappropriate measures of student learning.

This directly supports this study's findings, as 58.33% of teachers "Strongly Disagreed" that "standardized testing is an effective measure of student ability and learning" after the intervention. As previously mentioned, Fleming et al.'s quantitative data revealed concerns about the effectiveness of standardized testing, which was also a result in the qualitative teacher interviews, strengthening the overall finding. This is almost an identical finding to this study's theme of Standardized Testing & Curriculum Constraints, as participant 9 discussed how he "doesn't think there's anything in a traditional standardized assessment" that effectively demonstrates students "retention, absorption, and application . . . as long as standardized assessments are in play, then teachers are going to be hamstrung" (Interview 95:55-6:38). This similarity addresses the gap in the research, showing that the perceptions of Fleming et al.'s elementary school teachers aligns with the perceptions of the current study's high school teachers, which had yet to be studied.

Conclusions

Limitations

While this study offers valuable insights, limitations must be acknowledged to contextualize findings and guide future research. One limitation to take into account is the small sample size of 12 participants, which could make generalizing findings difficult for larger populations. Moreover, time constraints prevented the researcher from interviewing more participants, which limited the amount of data/findings. These time constraints resulted in an uneven distribution of STEM to Humanities participant ratio (8:4), which limited subpopulation comparisons and distinctions. Additionally, this research was conducted at Calabasas High School, which is a public school, meaning that these findings might only be relevant for the public educational sector. Due to these limitations, the researcher suggests future research to investigate private school perspectives in comparison.

Implications/Future Studies

As exemplified by the findings from this study, it is evident that teachers believe that IE implementation is feasible and beneficial to incorporate into ATE. For administrators and teachers, this study serves as a tangible example of how IE elements can be incorporated into ATE through pilot programs, staggered schedules, teacher involvement, etc. Drawing inspiration from Italian-style models, where student autonomy and experiential learning are central, these methods could offer a pathway to reform. For example, during Calabasas High School's 2023-2024 school year, high school stakeholders created a new bell schedule; however, staggered scheduling was not considered, as expressed by participant Two. Thus, in the future it should be a consideration based on the community's teacher perceptions and evidence-based benefits. To a further extent, the opinions of teachers on the weaknesses of ATE reveal how beneficial their perspectives are in the reformation process, as they are the voices of education within the classroom. Although this study only focused on teacher perspectives, for future

research it could be advantageous to understand student perspectives in the search for educational reform.

Conclusion

Through this study the researcher aimed to understand the most effective methods of implementing IE into the ATE high school system to address weaknesses within ATE through high school teacher perspectives. Using a mixed-method quasi experimental design the researcher was able to understand teacher perceptions qualitatively (semi-structured interview questions) and quantitatively (Likert-scale survey questions). Quantitatively, the researcher created Pre-Test versus Post-Test histograms, sub-population STEM v Humanities histograms, ran Pearson Correlation Coefficient Tests, and qualitatively, they conducted a thematic analysis on the 12 interviews.

Significant findings from this data analysis revealed that the most successful perceived implementation methods as described by the teachers are Scheduling/Classroom Innovations as well as Reform/Teacher Led Change. Within these identified themes, participants for both Humanities and STEM subjects discussed how pilot programs, schedule and classroom changes, and teacher involvement can facilitate this integration. These findings directly answer the research question as it provides viable options to integrating IE into ATE high schools, a previously unexplored area. Moreover, when asked the question “What do you see as the most significant challenges or weaknesses in ATE?” the majority of participants cited that large class sizes were the most significant weakness of the ATE system. This is due to large class sizes preventing teachers from addressing all students' needs individually and cultivating meaningful relationships. Therefore, contributing to lowered academic outcomes and engagement issues because of the “one size fits all” model in ATE. Hence, through identification of the negative impact of large class sizes in ATE, it addresses part of the research question discussing the weaknesses within the system.

On the quantitative side, the fact that participants chose lower categories to the Likert scale questions in the post-test illustrates how the intervention was successful in changing perceptions. The overall increase in negative perceptions towards ATE reveals that teachers have significant opposition to the traditional system within their high school. Previous studies, such as Martalock and Randolph et al., did not consider the high school or teacher demographic. In contrast, this research addresses that gap by exploring high school teachers' perspectives through interviews to understand a previously neglected group.

Works Cited

- Aljabreen, H. (2020). Montessori, waldorf, and reggio emilia: A comparative analysis of alternative models of early childhood education. *International Journal of Early Childhood*, 52(3), 337–353.
<https://link.springer.com/article/10.1007/s13158-020-00277-1>
- Chicoine, D. (2004). Ignoring the Obvious: A Constructivist Critique of a Traditional Teacher Education Program. *Educational Studies*, 36(3).
https://doi.org/10.1207/s15326993es3603_4
- Corrigan, D. (2013). Why a Comprehensive Approach to Educational Reform is Necessary: Children learn what they live outside of, as well as inside of schools. UVM ScholarWorks. <https://scholarworks.uvm.edu/jmjcpr/6/>
- Fleming, D. J., Culclasure, B. T., Warren, H., & Riga, G. (2024). The challenges and opportunities of implementing Montessori education in the public sector. *Journal of Montessori Research*, 10.
- Home. (2025). Lighthouse Academy of Fredericksburg. <https://kidslaf.com>
- Instructional Strategies | Centre for Teaching and Learning. (2025). Queensu.ca.
<https://www.queensu.ca/ctl/resources/instructors/instructional-strategies>
- Kannwischer, M. (2004). Development of the early childhood curricular beliefs inventory: an instrument to identify preservice teachers' early childhood curricular orientation (Doctoral dissertation, Dissertation, Florida State University).
- Kristiyani, C. (2018). MATERIALS AND (LANGUAGE) LEARNING ENVIRONMENT BASED ON MONTESSORI CONCEPTS. *Language and Language Teaching Journal*, 21(1), 46–54. <https://doi.org/10.24071/llt.2018.210105>
- Martalock, P. L. (2012). What is a wheel?" the image of the child: Traditional, project approach, and Reggio Emilia perspectives. *Dimensions of Early Childhood*, 40(3), 3-12.
- Minkin, R. (2024, April 4). About Half of Americans Say Public K-12 Education Is Going in the Wrong Direction. Pew Research Center; Pew Research Center.
<https://www.pewresearch.org/short-reads/2024/04/04/about-half-of-americans-say-public-k-12-education-is-going-in-the-wrong-direction/>
- Montessori vs. Traditional Education | American University. (2024, September 12). School of Education Online.
<https://soeonline.american.edu/blog/montessori-vs-traditional-education/>
- Moon, T., Brighton, C., Jarvis, J., & Hall, C. (2007). THE NATIONAL RESEARCH CENTER ON THE GIFTED AND TALENTED State Standardized Testing Programs: Their Effects on Teachers and Students. <https://files.eric.ed.gov/fulltext/ED505375.pdf>
- National Center for Montessori in the Public Sector (NCMPS). (n.d.). National Center for Montessori in the Public Sector. <https://www.public-montessori.org/>
- Oaks, M. (2024, June 4). A Day in the Life of a Montessori Student: Experience Learning Through Their Eyes. Montessorioaks.com.
<https://www.montessorioaks.com/a-day-in-the-life-of-a-montessori-student-experience-le>

arning-through-their-eyes

Office for Health Improvement and Disparities. (2021, September 8). Quasi-experimental study: Comparative studies. GOV.UK.

<https://www.gov.uk/guidance/quasi-experimental-study-comparative-studies>

Randolph, J., Bryson, A., Menon, L., Henderson, D. K., Manuel, A. K., Michaels, S., Leigh, D., McPherson, W., O'Grady, R., & Lillard, A. S. (2023). Montessori education's impact on academic and nonacademic outcomes: A systematic review. *Campbell Systematic Reviews*, 19(3). <https://doi.org/10.1002/cl2.1330>

Rathunde, K. (2003). A COMPARISON OF MONTESSORI AND TRADITIONAL MIDDLE SCHOOLS: MOTIVATION, QUALITY OF EXPERIENCE, AND SOCIAL CONTEXT. *The NAMTA Journal* •, 28(3). <https://www.countrysideday.org/docs/rathundecompar.pdf>

\Stegelin, D. A. (2003). Application of the Reggio Emilia approach to early childhood science curriculum. *Early Childhood Education Journal*, 30, 163-169.

The Effects of Music Genres on EEG Waves By Isabelle Jiang

An EEG, otherwise known as an electroencephalogram, is a type of test done to record the brain's electrical activity. These electrical signals come from neurons near the brain's surface when they fire in large groups around the same time to send information across synapses. However, a major issue associated with EEG is the contamination of the data from other electrical signals, such as from the body. Nevertheless, EEG is one of the best (and non-invasive ways) to record brain activity. There are 5 main types of waves:

Alpha: (α : 7.5-13 Hz), is commonly associated with a relaxed state of consciousness. Alpha waves are beneficial for reflective and reactive purposes.

Delta: (δ : 1-4 Hz), is most commonly associated with sleep, and specifically, NREM sleep. In a conscious state, can also be associated with feeling relaxed and the subconscious.

Theta: (θ : 4-8 Hz), is most commonly associated with memory. Theta waves are believed to originate mainly from the hippocampus.

Beta: (β : 13-30 Hz), is most commonly associated with a state of focus. A high amount of these waves are also associated with overstimulation and stress.

Gamma: (γ : 30-44 Hz), is most commonly associated with high attention. These waves are the least studied due to often being contaminated by other factors.

In this project, the 5 different waves were measured through four sensors: AF7, AF8, TP9, and TP10. In a clinical EEG, many more sensors are used. However, in this study, only four were used as a general overview of each part of the brain.

AF7 and AF8: Left and right sections of the prefrontal cortex, respectively. This section of the brain is mainly in charge of decision-making, social awareness, voluntary muscle movements, and the ability to learn as well as your memory.

TP9 and TP10: The left and right temporal lobes, respectively. This section of the brain plays a large role in memory (as the hippocampus is within it), the processing and usage of language (Wernicke's area is within it), emotional processing (the amygdala is within it), the processing of auditory information, and visual recognition.

Music therapy is defined as the usage of music (such as rhythm or sound) for the sake of well-being. This form of therapy is used by healthcare providers in a clinical setting, such as a therapy session. However, in this project, music therapy will be defined as listening to music to induce wellbeing (such as reducing stress.)

To do list

- create a consent form

Methods

Participants

In this preliminary investigation, we conducted a pilot study to assess the feasibility of using the MUSE EEG headband for measuring neural responses to different music genres. A small sample of 5 participants (3 females, 2 males) was recruited, including 2 musicians and 3 non-musicians, to test our experimental protocol and data collection methods.

Data Collection and Hardware Specifications

This pilot study utilized the MUSE headband to record neural activity through four electrodes (TP9, AF7, AF8, TP10) at a sampling rate of 256 Hz with a 60 Hz notch filter applied to remove power line interference. We collected absolute band powers (calculated from the logarithm of the Power Spectral Density of the EEG data), raw EEG signals (0.0-1682.815 μ V range), accelerometer data (gravity measurements from -2 to +2 g), and gyroscope measurements (-245 to +245 degrees/second) every second for a period of 20 seconds. To ensure data quality, we monitored the HSI (Horse Shoe Indicator) values for each sensor (1=Good, 2=Medium, 4=Bad) and the HeadBandOn indicator (1=True, 0=False), which helped us identify and address potential electrode placement issues during the testing phase.

Signal Processing and Analysis

We focused on analyzing five frequency bands: Delta (δ : 1-4 Hz), Theta (θ : 4-8 Hz), Alpha (α : 7.5-13 Hz), Beta (β : 13-30 Hz), and Gamma (γ : 30-44 Hz). The absolute band powers typically ranged from -1 to +1. We developed and tested a custom-built Python script for data processing and visualization. Our script plots and examines the frequency bands across the sensors.

Preliminary Protocol Testing

Each participant underwent a test session listening to three 20 second musical selections: Iron Maiden's "The Trooper" (heavy metal), Mozart's Symphony No. 40 in G Minor (classical), and a self-selected piece. This pilot phase allowed us to identify several key methodological considerations: optimal headband placement procedures, necessary duration of rest periods between musical selections, and potential confounds from movement artifacts. We tested a basic correlation analysis between wave types and sensor locations using Python's *seaborn* heatmap visualization, which showed promising potential for identifying relationships between brain activity patterns and musical stimuli. A prototype Google Forms questionnaire was also tested for gathering participant feedback and musical background information.

Results

Average Values by Wave Type and Condition				
	Control	Maiden	Mozart	Preferred
Alpha	0.641	0.638	0.671	0.636
Beta	0.294	0.254	0.244	0.214
Delta	0.537	0.429	0.303	0.329
Gamma	-0.068	-0.104	-0.083	-0.139
Theta	0.317	0.236	0.174	0.208

Figure 1: Average Values of Wave Type and Condition

The collected data included 20-second recordings of 5 participants for each song. Control recordings were taken and graphed alongside the data for comparison. Figure __ shows the average amplitude values of each wave for each song. This section will take a look at the comparison of each wave separately.

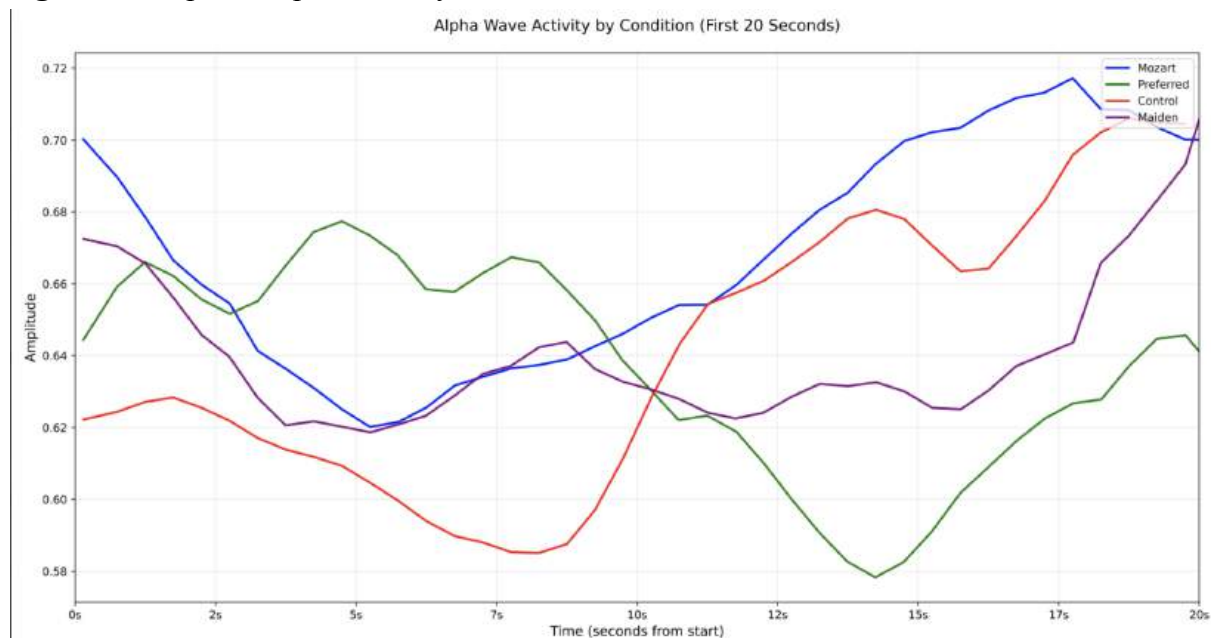
Alpha

Figure __: Alpha Wave Average Values by Song Type

Alpha Wave Average Values by Condition				
Condition	Average	Std Dev	Subjects	Files
Mozart	0.671	0.196	5	2
Preferred	0.636	0.211	5	3
Control	0.641	0.237	5	1
Maiden	0.638	0.197	5	1

As can be seen in Figure 1, Alpha waves were at their highest amplitude for Mozart and had the lowest standard deviation. Alpha waves were at their lowest while participants listened to their preferred music. The high standard deviation for Control and Preferred is understandable as participants either listened to different songs or simply had different wave activities.

Figure 2: Graph of Alpha Activity



Looking at this graph, we see a common positive pattern between all conditions. For each condition, the alpha waves showed a negative pattern in the beginning, but after the midpoint, began to increase in amplitude. Though the graph stops at 20 seconds, the continuation of this positive trend can be inferred.

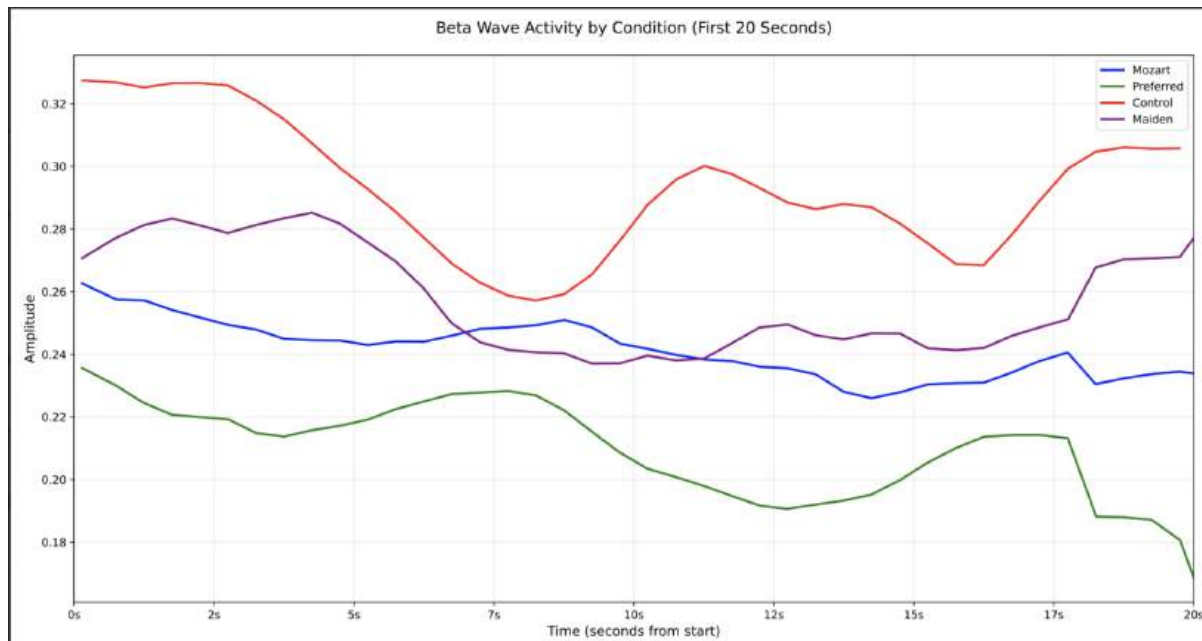
Beta

Figure __: Beta Wave Average Values by Song Type

Condition	Average	Std Dev	Subjects	Files
Mozart	0.244	0.147	5	2
Preferred	0.214	0.165	5	3
Control	0.294	0.189	5	1
Maiden	0.254	0.209	5	1

This figure shows the average amplitudes of Beta waves for the different conditions. Disregarding control, the Iron Maiden song caused the highest Beta waves, though with a large

standard deviation, suggesting that the values varied highly between the subjects. The lowest amplitude was the Preferred.

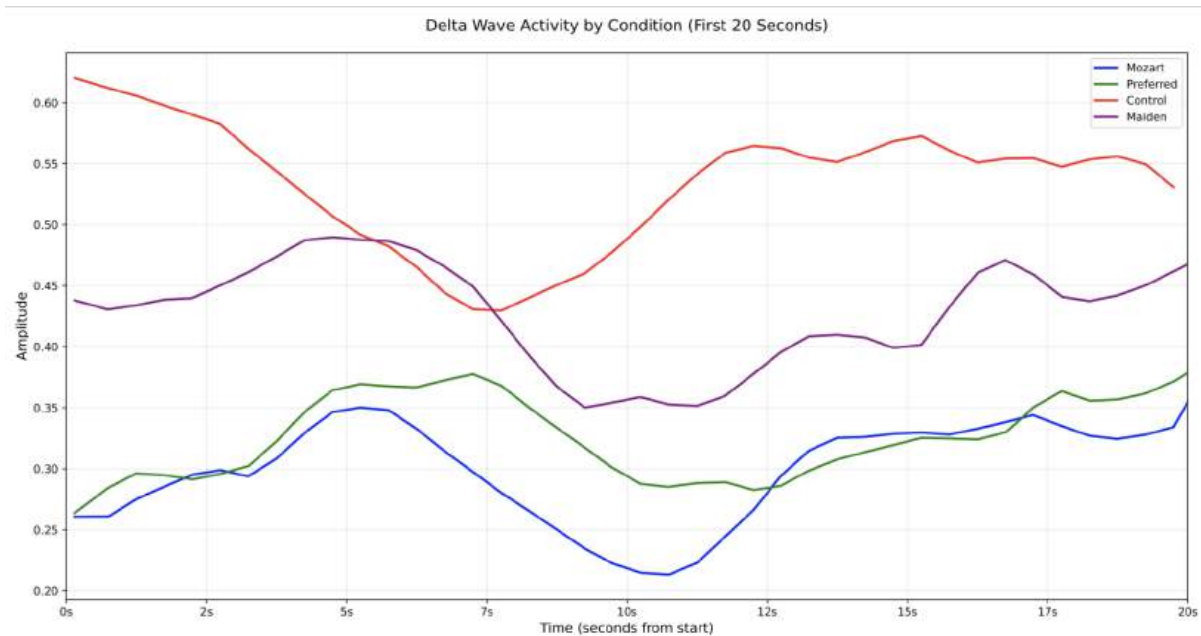


There isn't a clear pattern between the four conditions, but it is clear that the preferred music stayed much lower in amplitude compared to Iron Maiden and Mozart, which had a somewhat similar amplitude throughout.

Delta

Condition	Average	Std Dev	Subjects	Files
Mozart	0.303	0.301	5	2
Preferred	0.329	0.29	5	3
Control	0.537	0.411	5	1
Maiden	0.429	0.291	5	1

Looking at the table, the highest average amplitude (disregarding control) for Beta waves was the Iron Maiden song. The lowest average amplitude was recorded from Mozart, which also had the highest standard deviation, suggesting that the amplitude of Beta waves for each participant have a large difference.



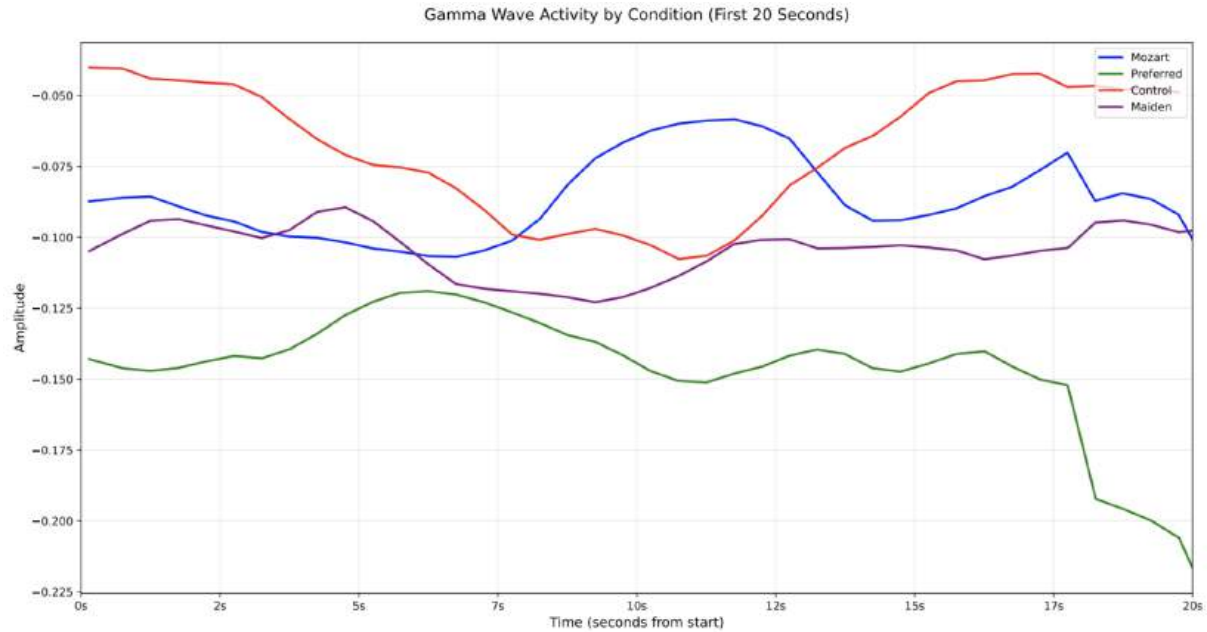
In this graph, we again see the Maiden waves above preferred songs and Mozart. Similar to previous gaps, we see a slight increase in the amplitude of all three conditions after the half-way point.

Gamma

Gamma Wave Average Values by Condition

Condition	Average	Std Dev	Subjects	Files
Mozart	-0.083	0.142	5	2
Preferred	-0.139	0.185	5	3
Control	-0.068	0.184	5	1
Maiden	-0.104	0.223	5	1

For gamma wavelengths, all amplitudes were in the negatives. The highest average was seen for Mozart, and the lowest was seen in the preferred. Compared to control, the biggest change was seen in the preferred condition.



Compared to other graphs, less of an overall pattern is noticeable in Gamma wave activity. There isn't a single condition that stayed above the others, but the Gamma wave activity for the Preferred condition consistently stayed below the others. Additionally, the Preferred conditions continued in a negative slope whereas the others stayed roughly the same.

Discussions

For a state of well-being, we will be looking at a state of relaxation to combat stress. Thus, the waves that should be brought into focus are Alpha and Delta. Looking at Alpha wave activity from the table, the highest amplitude of Alpha was seen in Mozart, suggesting relaxing qualities seen within classical music. However, looking deeper into the song, we see that though this is a classical piece, a strong amount of energy is associated with it, especially within the first 20 seconds. Thus, a classical piece with a high tempo would be the most beneficial for relaxation. Looking at Delta activity, the highest amplitude (excluding control) was seen in the Maiden. Both these songs have an energetic feel in common, though their differences lie within the instruments. Ultimately, this data shows that the best way to use music therapy to relax is to listen to energetic and unheard-before songs.

Costs and Care for AD/ADRD Patients: Quantitative Analysis of Insurance Premium and Quality of AD/ADRD Treatment (2017-2021) By Kaavya Veligeti

I. Abstract

On a global scale, neurodegenerative diseases are one of the most prevalent health conditions, yet due to their complexities, patients across the world continue to struggle with lack of quality treatment. In 2020, the World Health Organization determined that there were greater than 55 million people affected by AD/ADRD globally.⁴⁵ As described by Tay et al., Professor of Economics from Stanford University, there is an immense economic burden of AD/ADRD care, with costs of up to \$171,283.80 per capita in severe cases.⁴⁶ Globally, the costs associated with AD/ADRD are projected to reach \$14.5 trillion by 2050 with a greater burden on lower-income populations.⁴⁷ In the United States especially, the cost of care is excessive with limited coverage by insurance agencies. Skaria, Doctor of Pharmacy from the University of Philadelphia, found that caregivers of AD/ADRD patients pay around \$12,388 out-of-pocket, making up around 25% of the cost of formal care.⁴⁸ In recent years, research into AD/ADRD has been accelerating, with most contemporary literature consisting only of examinations of individual metrics (functional outcomes, healthcare utilization, or patient satisfaction) for AD/ADRD in relation to cost or quality of care. However, no literature has yet studied these metrics in association over a time period significant enough to establish a nuanced, holistic understanding of their relation. Therefore, no literature has yet explored a direct relationship between cost of Medicare premium (CMP) and quality of treatment. The goal of this research is to answer the question: To what extent is there a statistically significant correlation between cost of Medicare premium and quality of treatment for AD/ADRD patients in the United States over the time period of 2017-2021? A quantitative data fusion method was used to establish a comprehensive definition of treatment quality which was then used to determine the extent of its relationship with cost. This was done by fusing multiple raw datasets into one dataset used to perform a correlational analysis. This analysis can be used as a determination of factors that definitively impact quality of care for AD/ADRD populations in the United States in order to understand what future steps should be taken to improve quality of treatment, and overall quality of life for affected individuals.

⁴⁵ “ADI - Dementia Statistics,” ADI - Dementia Statistics, n.d., <https://www.alzint.org/about/dementia-facts-figures/dementia-statistics/>.

⁴⁶ Lyn Xuan Tay et al., “Economic Burden of Alzheimer’s Disease: A Systematic Review,” *Value in Health Regional Issues* 40 (November 14, 2023): 1–12, <https://doi.org/10.1016/j.vhri.2023.09.008>.

⁴⁷ Amy Lastuka et al., “Societal Costs of Dementia: 204 Countries, 2000–2019,” *Journal of Alzheimer S Disease* 101, no. 1 (August 12, 2024): 277–92, <https://doi.org/10.3233/jad-240163>.

⁴⁸ Anita Pothen Skaria, “The Economic and Societal Burden of Alzheimer Disease: Managed Care Considerations,” *The American Journal of Managed Care* 28, no. Suppl 10 (September 1, 2022): S188–96, <https://doi.org/10.37765/ajmc.2022.89236>.

II. Literature Review

Profit-Based Healthcare and Cost of Care in Aging Populations

In recent decades, the American healthcare system has shifted to a profit-based system based on a fee-for-service model. This model has led to an increasing prioritization of profits in managed care in the United States, which can inhibit access to healthcare due to threats of high medical bills, according to Christensen, Professor of Sociology at Midwestern State University.⁴⁹ Similarly, as described by Carrol et al., Director of Health Economics at University of Sheffield, this phenomenon is especially prevalent in aging populations due to lower expression of factors affecting equity of access, such as “capacity for decision-making,” “awareness,” and “personal, social, and cultural circumstances.”⁵⁰ This handicap is further amplified in the case of patients with AD/ADRD because a large amount of long-term treatment is not covered by insurance agencies. As described by Hudomiet et al., Doctor of Economics from University of Michigan, living with AD/ADRD for even a short period can significantly increase lifetime costs, thus intensifying the cost barriers described by Christensen and Carrol.⁵¹ Building on this point, Chandra et al. advance concerns for the lack of insurance coverage potentially causing worsened health outcomes for these patients, regardless of the excessive costs they may be paying. Chandra et al., describes how “the average annual premium for an individual policy was \$2,700,” and generally only covers around \$150 per day for up to four years; however, “nursing home care typically costs upwards of \$100,000 per year,” so patients often must choose between paying excessive fees or to forgo necessary treatment.⁵² Overall, a literary consensus is that AD/ADRD coverage is significantly less than what is needed relative to disease severity.

Complexity of AD/ADRD Treatment

AD/ADRD is a classification for highly individualized neurodegenerative diseases that are characterized by memory loss, cognitive decline, and worsened functional outcomes, resulting in crucial complexities relating to diagnosis and treatment. As described by The National Institute of Neurological Disorders and Stroke, AD/ADRD is a disease which is “devastating” for affected individuals and is caused by a combination of “genetic, lifestyle, and environmental factors” which makes it highly individualized and hard to treat.⁵³ Furthering the

⁴⁹ Isaac Christiansen, “Commodification of Healthcare and Its Consequences,” *World Review of Political Economy* 8, no. 1 (April 1, 2017), <https://doi.org/10.13169/worlrevipoliecon.8.1.0082>.

⁵⁰ Christopher Carroll et al., “Equity in Healthcare Access and Service Coverage for Older People: A Scoping Review of the Conceptual Literature,” *Integrated Healthcare Journal* 4, no. 1 (January 1, 2022), <https://doi.org/10.1136/ihj-2021-000092>.

⁵¹ Péter Hudomiet, Michael D. Hurd, and Susann Rohwedder, “The Relationship Between Lifetime Out-of-pocket Medical Expenditures, Dementia, and Socioeconomic Status in the U.S,” *The Journal of the Economics of Ageing* 14 (November 30, 2018): 100181, <https://doi.org/10.1016/j.jeoa.2018.11.006>.

⁵² Amitabh Chandra, Courtney Coile, and Corina Mommaerts, “What Can Economics Say About Alzheimer’s Disease?,” *Journal of Economic Literature* 61, no. 2 (June 1, 2023): 428–70, <https://doi.org/10.1257/jel.20211660>.

⁵³ NINDS, “Progress in Alzheimer’s Disease and Related Dementias (AD/ADRD) Research,” National Institute of Neurological Disorders and Stroke, November 29, 2024, <https://www.ninds.nih.gov/news-events/directors-messages/all-directors-messages/progress-alzheimers-disease-and-related-dementias-adardr-research>.

complexity, the disease is often present for decades before the symptoms become apparent enough to seek aid, therefore inhibiting identification and treatment options. Resultantly, the majority of treatment options can only help retain quality of life rather than provide a cure. As a whole, this drastically increases the costs of treatment and puts greater strain on both the patient and caregiver as they attempt to gain desperately needed care. These characteristics of AD/ADRD cause severe mental, emotional, and financial burdens on the patient and caregivers. Overall, AD/ADRD is a complex disease to treat due to its individualized nature, yet the severity of the disease in combination with a lack of sufficient treatment options means that there is a heavy strain associated with AD/ADRD.

Factors affecting AD/ADRD Quality of Treatment

The three key factors affecting AD/ADRD quality of treatment, as most commonly described throughout literature, are: functional outcomes, healthcare utilization, and patient satisfaction. The functional outcomes of patients with AD/ADRD include their ability to perform activities of daily living (ADLs), cognitive abilities, and social abilities. As described by Maresova et al., ability to perform ADLs decreases as disease severity increases.⁵⁴ While the meta-analysis conducted by Maresova et al. used a small sample size of only 12 total sources, this study suggests a relationship between functional outcomes and dementia severity, demonstrating a greater reliance on formal care or informal care. There is a general literary consensus that formal care is the most effective treatment option for AD/ADRD, because it provides quality care that can allow a patient to retain their quality of life despite disease progression. The most prevalent form of formal care is nursing home care, therefore it constitutes the majority of healthcare utilization for AD/ADRD. Accordingly, a longitudinal study on the costs of various treatments related to AD/ADRD by Fishman et al., Professor of Health Services at the University of Washington, found that the cost of nursing home care was the largest percentage of financial burden and healthcare utilization.⁵⁵ The findings of Fishman demonstrate how dementia progression causes worsened functional outcomes while amplifying the costs of care and its associated financial barriers. The most complex metric to study is patient satisfaction as it is highly subjective and therefore difficult to quantify. Thus, literature attempting to study satisfaction has reported varying results. For example, Fenton et al., researcher at the University of California Center for Healthcare Policy, conducted a prospective-cohort study which determined a minimal correlation between satisfaction and costs of care,⁵⁶ however another study conducted by Park et al. found no relation.⁵⁷ While the Park et al., study may not be entirely

⁵⁴ Maresova et al., “Activities of Daily Living and Associated Costs in the Most Widespread Neurodegenerative Diseases: A Systematic Review”

⁵⁵ Bailey C Ingraham et al., “Caregiving for Dementia: Trends Pre-post Onset and Predictive Factors of Family Caregiving (2002–2018),” *Health Affairs Scholar* 2, no. 3 (February 16, 2024), <https://doi.org/10.1093/haschl/qxae020>.

⁵⁶ Joshua J. Fenton, “The Cost of Satisfaction,” *Archives of Internal Medicine* 172, no. 5 (February 14, 2012): 405, <https://doi.org/10.1001/archinternmed.2011.1662>.

⁵⁷ Park et al., “Public Satisfaction With the Healthcare System Performance in South Korea: Universal Healthcare System,” January 22, 2016.

accurate as it only included two agencies, it still demonstrates an opposing finding which shows the subjectivity of relationships with satisfaction. Overall, prior literature has investigated the correlation between each factor and costs, however there is still a crucial gap in determining their collective impact on the expenses of AD/ADRD treatment.

Considering the critical gap identified, my research aims to fill the gap by exploring various metrics for quality of treatment and unifying them into a singular dataset to be compared to insurance premiums over 5 years. This time period was determined due to the accessibility and uniformity of Medicare data across those years. As Medicare is the most commonly used insurance agency by Americans over the age of 65, which is the primary age group affected by AD/ADRD, using this insurance agency for data will provide a larger sample size and therefore more accurate results. My study will look at values for functional outcomes, healthcare utilization, and patient satisfaction for Medicare-insured AD/ADRD patients, in order to gain a holistic, nuanced understanding of quality of treatment. Understanding a potential correlation between these values will help determine if higher costs actually correlate to greater treatment quality or if these costs serve as a barrier to necessary treatment. This understanding will provide insights into the overall socioeconomic impact of insurance costs, which is especially important for aging populations who already face healthcare disparities and are affected by a complex disease with minimal treatment options.

III. Methodology

Process

This research began with the general topic of Alzheimer's Disease healthcare treatment and the impact of insurance. With sufficient research into variables affecting care and the lack of research into specific correlations of quality and costs over time, this topic evolved into the question: To what extent is there a statistically significant correlation between cost of Medicare premium and quality of treatment for AD/ADRD patients in the United States over the time period of 2017-2021? To determine a statistical correlation, the methodology used is data fusion to conduct a correlational analysis. I evaluated other literature^{12,13,17} in the field of health economics, most notably those attempting to identify correlation or causation between two or more variables, and I found that a correlational analysis was the most widely used and would be the most accurate method to fulfill the goal of this research. This is a quantitative methodology which uses statistical tests to identify potential correlations. From there, I began to more closely look at the ways I could obtain quantitative data relating to my topic; upon further literary exploration, I determined that data fusion was the best suited for my research. Data fusion combines data from a variety of sources into a singular dataset. As the factors for quality of treatment have been identified by literature¹⁸ but no one has yet studied quality as a singular factor, data fusion would allow me to combine datasets for these factors into a singular dataset which would fill the gap I previously identified. As Medicare is the most widely used health insurance agency by Americans who are most likely to be affected by AD/ADRD, I used the

Center for Medicare Services (CMS) to gather quantitative data for CMP and for the metrics of quality of treatment. Using Medicare as a source for both datasets reduces the impact of outside variables influencing costs, however, limiting this dataset to only Medicare may also limit the generalizability of these findings. To acquire a dataset for CMP, I found the values for the CMP for each year on the CMS website. To acquire a dataset for quality of treatment I used the Medicare Current Beneficiaries Survey (MCBS) and Health Outcomes Survey (HOS) which are Medicare Public Use Files (PUFs). These are publicly accessible files from the CMS, so the data they contain is entirely confidential. These files are sources of raw, quantitative data over the years 2017-2021 which have consistent questions across the time period, therefore allowing a more accurate understanding of trends. Then, I classified the questions from these surveys into the three metrics (functional outcomes, utilization, and satisfaction) relevant to quality of treatment for AD/ADRD. Next, the data was 'cleaned,' which essentially means checking for missing data, inconsistencies, potential errors, and getting rid of missing fields. Originally, this study was planned to be conducted using data from 2012-2021; however, due to significant inconsistencies and missing data prior to 2017, it was necessary to adjust the time period to 2017-2021 after cleaning. As this research was conducted for five years of data and two sets of files, there were a total of 10 files used, each containing around 200 survey questions with around 12,000 responses each. While the large dataset was beneficial in determining a statistical correlation, it was too immense to attempt to manually aggregate the data, so it was necessary to design a coding framework. By again reviewing other literature^{13, 20} in this field, I found that the Python software was the most commonly used for quantitative correlational analyses, so with aid from the Pandas Dataframe Assistance software, I was able to design a framework to fuse this data. It is important to note that using a coding framework may introduce aggregation bias if the variables are improperly scaled, however this research used a standard scaling system (0-1) in order to minimize this bias. Fusing this data was done by uploading the files into the framework and selecting the specific fields that should be analyzed for each metric, using the software to scale the values for each response from 0-1 (in which a 0 indicates this feature is not present and a 1 indicates that this feature is completely present at a high level), then average them into a single value to provide the composite score for a single metric. This process was repeated for each metric for each year, creating a single composite score for each metric for each year from 2017 to 2021. Then, the metrics were compressed into a single value for each year, thus providing a quantitative dataset for quality of treatment. Overall, by using data fusion, this research was able to determine a holistic, quantitative understanding of quality of treatment while streamlining the overall research process and reducing errors due to the use of software to aggregate the data. Finally, the correlational analysis is conducted using statistical tests to determine the extent of the correlation between quality of treatment and CMP. From my prior knowledge of statistics, I determined the tests necessary for this research to be a Least-Square-Regression Line (LSRL) test, a Linear-Regression T-Interval, a Quadratic-Regression Model, a Cubic-Regression Model, and a residual test. All of these tests are conducted by graphing the datasets and using a graphing calculator to run the tests. The two

tests conducted to determine the extent of a correlation are a Least-Square-Regression Line (LSRL) test to determine linear correlation strength by providing an r-value, in which a value over 0.85 correlates to a significant correlation; and a Linear-Regression T-Interval, in which an interval that does not contain zero indicates confidence in a correlation and the sign of an interval determines the slope of the correlation. Next, to further explore a potential correlation I used a Quadratic-Regression Model, which provides an r-value for a quadratic relationship between the values, a Cubic-Regression Model which provides a dataset for a cubic relation between the values, and a residual test which calculates the difference between the best-fit-line and the dataset. This methodology and the holistic measurement of quality additionally can provide a basis for future research into patient experiences and costs AD/ADRD.

Ethical Consideration

My research question was sent into an Institutional Review Board in order to ensure compliance with ethical standards. After approval, I was able to proceed with the research. This research was conducted using Medicare PUFs, specifically the HOS and MCBS, from the CMS. It is necessary to ensure that all patient privacy is protected to comply with HIPAA regulations, which is done by using de-identified data, and complying with the CMS PUF's User Agreement, which states that data for research must be used with integrity for research purpose. This research meets these requirements by using complete datasets to ensure an accurate understanding. All PUFs contain raw data with randomized patient numbers, and completely de-identified raw datasets, meaning there are no direct identifiers (name, address, phone number, social security number, etc.) Indirect identifiers (age, race, gender, etc.) were not factors for these metrics, therefore they are not present in this research. The data was then aggregated into a dataset of composite values so there is no possibility of identifying patient information from the data in this research paper.

Limitations

The limitations of the methodology are in relation to the time constraints and the data source. Medicare was the only insurance agency used because it has publicly available data so there was limited access to data. The short amount of time available in which I would be able to conduct the study additionally aided the decision to only use one agency, as it would not be feasible to conduct data integration for multiple different insurance agencies within the given time frame. Additionally, due to the incomplete data prior to 2017, it was necessary to decrease the time period of the research from 2012-2021 to 2017-2021. Furthermore, a limitation of the data gathered is that it excludes the impact of external influences. While quality of treatment is heavily affected by the determined metrics, to a lesser extent it is also impacted by varying factors such as regional healthcare access, provider expertise, and long-term care availability. However, as this study was focused specifically on determining values for quality of treatment and evaluating correlation with CMP, the influence of other variables are outside the scope of this research as that would alter the independent analysis of the relationship between the two

chosen datasets. Furthermore, the chosen methodology of data fusion is designed to limit external influences due to the inclusion of multiple variables which account for many potentially confounding variables in order to evaluate correlation.

IV. Results

Upon an initial look at the values presented in Fig 1., the Quality Score has minimal increases at an average rate of +0.00047 points per year (PPY). On the other hand, CMP increases more significantly, with an initial value of \$134 USD and a final value of \$148.50 USD. In more closely examining the changes in three metrics, the data reveals that satisfaction (+0.00289 PPY) and functional outcomes (+0.00079 PPY) increase while utilization decreases (-0.00228 PPY). This finding is particularly unexpected as a primary aim of health insurance is to aid utilization, and as described by much literature, including the meta-analysis by University of California Doctor of Economics E. Shami⁵⁸ and the policy analysis by Economist A. Hardiman⁵⁹ at the American Medical Association, utilization is expected to increase to a proportional to insurance costs. It should be noted that any interpretations of this finding should be tentative of potential external factors.

Year	Cost of Medicare Premium (CMP)	Functional Score	Utilization Score	Satisfaction Score	Quality Score
2017	\$134.00	0.79820	0.49595	0.76680	0.68698
2018	\$134.00	0.79744	0.49527	0.76669	0.68647
2019	\$135.50	0.79896	0.49221	0.77165	0.68761
2020	\$144.60	0.79988	0.48969	0.77579	0.68846
2021	\$148.50	0.80093	0.48736	0.77672	0.68834
Rate	\$3.96/year	+0.00079/year	-0.00228/year	+0.00289/year	+0.00047/year

Figure 1. Year-By-Year Cost of Medicare Premium, Metrics of Quality, and Quality of Treatment.

Then, the values for quality were compared to the costs in a graph in order to display the trends. A preliminary visual analysis of Fig. 2 shows a relatively strong correlation between CMP and quality which will be further explained in the analysis section. An unusual finding, however, is that the graphs appear to diverge in 2021.

⁵⁸ Elham Shami, Jafar Sadegh Tabrizi, and Shirin Nosratnejad, "The Effect of Health Insurance on the Utilization of Health Services: A Systematic Review and Meta-Analysis," *Galen Medical Journal* 8 (October 14, 2019): 1411, <https://doi.org/10.31661/gmj.v8i0.1411>.

⁵⁹ Allen Hardiman et al., "National Health Expenditures, 2023: Health Care Spending Growth Outpaced GDP for the First Time Since the Pandemic as Utilization Rose," American Medical Association, 2025, <https://www.ama-assn.org/system/files/prp-health-care-spending-outpaced-gdp.pdf>.

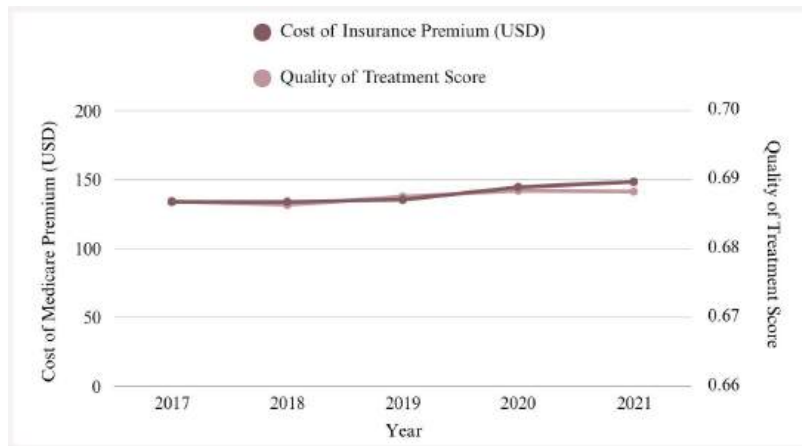


Figure 2. Cost of Medicare vs. Time & Quality of Treatment vs Time.

Various statistical tests were used to identify and evaluate a potential correlation between quality and CMP. The LSRL test found a resultant r-value of 0.886. Based on this value it is clear that values for quality increase proportionally to CMP; this means that there is a statistically significant correlation between these variables. The Linear Regression T-Interval found a positive interval at a 95% confidence level, meaning that there is a statistically significant positive correlation between CMP and Quality of Treatment.

To more thoroughly explore this finding, a test model was run using a quadratic regression and a cubic regression. The quadratic regression resulted in an r-value of 0.910. Interestingly, the cubic regression had the most significant r-value of 0.955, making it the best fit line for the data. To explore potential other statistical correlations, a residual plot was used with the result being no observable pattern, therefore the LSRL line is fit to describe the data accurately.

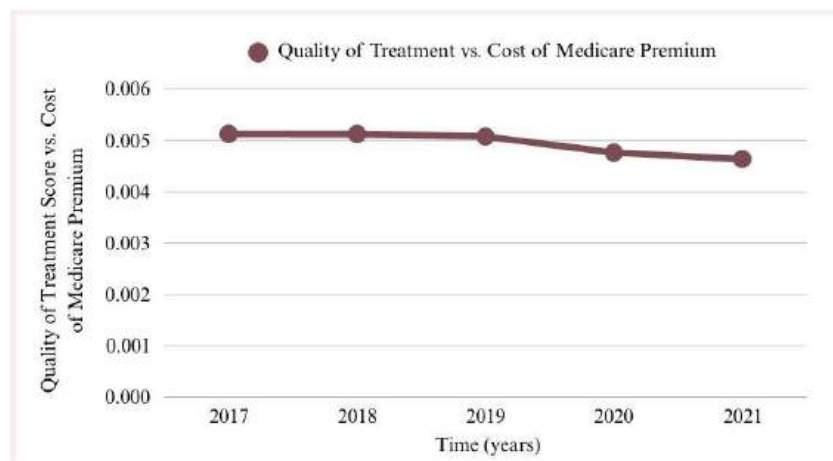


Figure 3. Quality of Treatment vs. Cost of Medicare Premium

Finally, the data was displayed on a graph showing the quality score per dollar of CMP per year, as depicted in the graph above (Fig. 3). The quality per dollar CMP appears to decrease over time. Interestingly, the most significant change in this relationship occurred in between

2019 to 2020-- a drop from 0.00507 points per dollar CMP to 0.00476 points per dollar CMP. This divergence highlights the necessity of policy evaluations and the impact of insurance costs on access and affordability of care.

V. Analysis and Discussion

Based on the results of the LSRL test and the Linear Regression T-Interval, there is a statistically significant positive correlation between Quality of Treatment for AD/ADRD patients and Cost of Medicare Premium. This correlation means that with CMP increases, quality of treatment is expected to increase as well. When evaluating the graph of Fig. 2, this correlation appears very strong during the years of 2017 to 2019. Then, in 2020, the lines for the graphs begin to diverge, with a clear separation by 2021. This is further established by the graph of Fig. 3 in which the correlation remains relatively consistent from 2017 to 2019, then there is a sharp drop in 2020 with consistent values for 2021. As previously stated, one of the key limitations of this research is that the variables used to determine a correlation does not account for the impact of external factors such as regional healthcare differences, inflation, socioeconomic differences, or a changing legal environment; thus, this decrease in correlation is most likely the result of external factors.

The change from 2019 to 2020 can be most directly attributed to the COVID-19 pandemic which widely altered the healthcare environment across the both in the United States and throughout the world, overall decreasing access to affordable high-quality treatment and decreasing overall health outcomes while dramatically increasing costs across numerous aspects of the healthcare system. As this research specifically evaluates the quality of treatment for AD/ADRD patients, it is plausible to hypothesize that the COVID-19 pandemic had an impact on the overall quality of treatment for AD/ADRD in the United States. While the coincidence of these events suggests causation, due to the methodology being correlational, causation cannot be established.

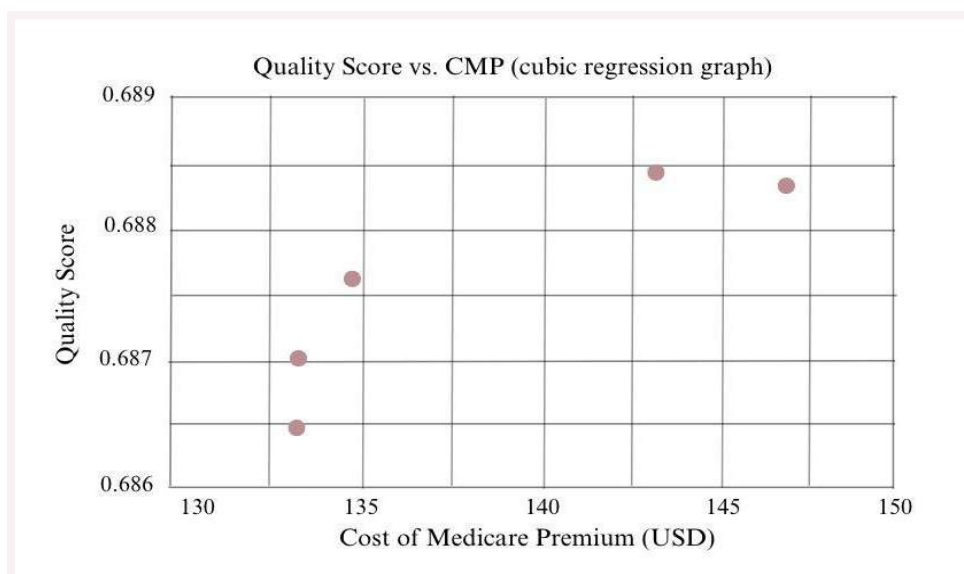


Figure 4. Cubic Regression Graph for Quality Score vs. CMP

This phenomenon can further be seen through the results of the cubic regression model. As there was a stronger correlation for a cubic model than for a linear model, the CMP increases at a slightly greater rate relative to quality of treatment as the rate of change for quality of treatment remains consistent over time, whereas the rate of change for CMP increases over time. This can be explained by looking at the curve in the cubic regression graph of Quality of Treatment vs CMP (Fig. 4), the curve in the graph decreasing over time as cost increases at a high rate, while quality of treatment more minimally increases at a slow rate. In recent years the costs have continued to increase, according to the U.S. Centers for Medicare & Medicaid Services (CMS), with a CMP of \$174.00 in 2024.⁶⁰ This is a significant increase from the costs in 2021, however the Medicare PUFs containing data from 2022 to 2024 have not yet been released, so quality of treatment for these years cannot be determined. Regardless, it can be hypothesized that the quality will continue to increase at a relatively constant rate with disproportionately large increases in CMP, which may decrease the strength of the linear correlation. This could potentially strengthen the correlation for the cubic graph rather than the linear graph, or it may even decrease any correlation at all between the two variables.

The confidence interval revealing a 95% confidence in a statistically significant positive correlation demonstrates that during this time period, quality and CMP increase with a directly proportional relationship. The correlation with a 95% confidence specifically reveals that these results are widely applicable on a greater population level with less uncertainty. Since there is only 5% of uncertainty for these results, it can be assumed that external factors do significantly influence the correlation. This further supports the methodological choice for this research to not attempt to factor in variance based on external stressor. Due to the low uncertainty percentage, the relatively minimal impact of these factors do not affect the relationship between quality and CMP enough to be able to confidently determine that external factors would alter the outcome, therefore excluding their impact in calculations has a negligible impact on the validity of the research as a whole.

Additionally, when exploring the trends for the metrics of quality of treatment as well as quality of treatment as a whole (Fig. 1), it should be noted while quality as a whole increases during the overall time period of 2017-2021, quality actually began to decline from 2020-2021. From 2017-2020, quality of treatment increases a total of 0.00148 points (0.68698 to 0.68846). However, from 2020 to 2021, there is a decline of 0.00012 points (0.68846 to 0.68834). This is a notable finding, as prior to 2020 quality was increasing at a rate of 0.00049 PPY, then 2020 marked a decrease to a rate of 0.00012 PPY. This suggests a shift to a trend of decreasing quality. This notion can be considered speculative since the publicly available datasets only range until 2021, therefore it is not clear if this decrease in 2021 is the start of a new trend, or if it would better present as merely an outlier in a larger dataset.

VI. Conclusion

⁶⁰ CMS. "Home - Centers for Medicare & Medicaid Services | CMS," n.d., <https://www.cms.gov/>.

The purpose of this research is to answer the question: To what extent is there a statistically significant correlation between cost of Medicare premium and quality of treatment for AD/ADRD patients in the United States over the time period of 2017-2021? From the analyses, there is a statistically significant positive correlation between quality and CMP, based on the LSRL r-value of 0.886 which demonstrates a strong correlation, and a positive confidence interval at a 95% confidence level which demonstrates high certainty. Another notable finding is the stronger correlation found using a cubic regression model (r-value=0.955). As seen in Fig. 4, this model indicates a plateau and slight decrease in quality near the end of the time period, therefore hinting at potentially weaker correlations in the years following 2021. The findings of this research align with the initial hypothesis for this study, which was that there will be a significant demonstrated correlation, although the cubic regression model adds a layer of complexity by suggesting changing trends.

Implications

Overall, the significant positive correlation suggests that increases in Medicare premium correlate to a higher quality of treatment. This raises concerns based on socioeconomic disparities between patients, as many patients have limited access to quality treatment due to cost barriers. Furthermore, the decreasing correlation, as presented in Fig. 4, suggests that it is necessary to re-evaluate the resource allocation of the funds obtained through CMP. It can be predicted that increasing costs would correlate to increasing quality because there is greater availability of money to spend on various treatments and care, so the decreasing relationship implies that resources may not be used in the most effective ways to support patient needs. Therefore, this suggests that it may be necessary for policymakers to evaluate the allocation of Medicare funds in order to maximize patient quality with respect to expenses.

The most unexpected result of this research was determined through a cubic regression model, as shown in Fig. 3, where the quality plateaus and then decreases at the very end of the time period evaluated. As this occurred at the end of the time period, conclusive interpretations about this event cannot be established because there is no future data to support any projected trends. Regardless, this data suggests two potential interpretations. The first is that while cost increases correlate to quality increases, this relationship has a maximum capacity, after which it begins to decline. This is suggested by the decrease in quality at the end of the time period while cost continues to increase which can be interpreted to mean that there may be an ideal CMP, after which subsequent increases instead correlate to decreasing quality. This implies that the ideal CMP occurred between 2017-2021. Therefore, the results of this research could be used to aid policymakers in evaluating prior policies to determine what most effectively promoted treatment quality for AD/ADRD patients in order to support future decisions. It must be noted, however, this outcome is hypothetical as this lies outside the time period, and it is based on a slight shift in a trend over the course of only two years. Another potential outcome is that the decrease in values for quality, and the resultant cubic correlation strength, are a factor of external influences. Specifically, this could be attributed to the COVID-19 pandemic which disproportionately

affected AD/ADRD patients because AD/ADRD is a neurodegenerative disease which negatively affects all functions, therefore these patients would be more significantly impacted by a novel disease with very minimal treatment options. While this outcome can be considered speculative because of the location of this datapoint at the end the time period, this outcome would suggest that since COVID was the disruption, the overall correlation remains strong and therefore policymakers would now be able to understand that quality must be maximized with respect to the costs of insurance premium, therefore resource allocation must also be re-evaluated. Regardless, this does provide insights into potential policy decisions that may be taken and factors that affect quality of treatment which can be used to aid future steps into ensuring equity of access and quality for treatment of AD/ADRD patients across the United States.

Future Directions

This study aimed to determine the correlation between the two datasets of quality of AD/ADRD treatment and CMP. Thus, this research excluded external factors so it may be beneficial for future research to evaluate the impact of these factors on the relationship established in this research to determine which factors most directly harm this relationship, and therefore which steps should be taken to aid greater overall health and treatment for AD/ADRD patients. Furthermore, this study was conducted from 2017-2021, yet a notable data point for quality of treatment was found at the end of the time period in 2021. Therefore, no conclusive interpretation could be developed. Additionally, in accordance with the methodology it was not possible for this study to evaluate data from years following 2021 because the Medicare PUFs for the years after 2021 have not yet been released. Thus, future research should conduct another longitudinal study including at least five years past 2021 in order to determine if this decrease is an outlier in the dataset caused by external factors, or if this signifies a shift towards a lasting trend resulting from excessive cost increases with quality of treatment not increasing fast enough to compensate. Another limitation of this research was that, due to time constraints, only one insurance agency was evaluated. Future research may want to explore this correlation using other insurers, both public and private, to determine if this trend is widespread or an event isolated to Medicare. Finally, this study was able to establish a statistically significant correlation between the values for quality and CMP, however this merely states a relationship. Future research should continue to explore this relationship in order to determine if these values are reliant on another external factor, or if there is a potential causation relationship between cost of Medicare premium and quality of AD/ADRD treatment.

Works Cited

- ADI - Dementia Statistics. "ADI - Dementia Statistics," (n.d.)
<https://www.alzint.org/about/dementia-facts-figures/dementia-statistics/>.
- Carroll, Christopher, Katie Sworn, Andrew Booth, Aki Tsuchiya, Michelle Maden, and Megumi Rosenberg. "Equity in Healthcare Access and Service Coverage for Older People: A Scoping Review of Conceptual Literature." *Integrated Healthcare Journal* 4, no. 1 (January 1, 2022). <https://doi.org/10.1136/ihj-2021-000092>.
- Chandra, Amitabh, Courtney Coile, and Corina Mommaerts. "What Can Economics Say About Alzheimer's Disease?" *Journal of Economic Literature* 61, no. 2 (June 1, 2023): 428–70. <https://doi.org/10.1257/jel.20211660>.
- Christiansen, Isaac. "Commodification of Healthcare and Its Consequences." *World Review of Political Economy* 8, no. 1 (April 1, 2017).
<https://doi.org/10.13169/worldrevipoliecon.8.1.0082>.
- CMS. "Home - Centers for Medicare & Medicaid Services | CMS," n.d. <https://www.cms.gov/>.
- Dauphinot, Virginie, Michele Potashman, Mihaela Levitchi-Benea, Ray Su, Ivana Rubino, and Pierre Krolak-Salmon. "Economic and Caregiver Impact of Alzheimer's Disease Across the Disease Spectrum: A Cohort Study." *Alzheimer's Research & Therapy* 14, no. 1 (February 12, 2022). <https://doi.org/10.1186/s13195-022-00969-x>.
- Fenton, Joshua J. "The Cost of Satisfaction." *Archives of Internal Medicine* 172, no. 5 (February 14, 2012): 405. <https://doi.org/10.1001/archinternmed.2011.1662>.
- Hardiman, Allen, American Medical Association, Centers for Medicare & Medicaid Services, and Et Al. Martin. "National Health Expenditures, 2023: Health Care Spending Growth Outpaced GDP for the First Time Since the Pandemic as Utilization Rose." American Medical Association, 2025. (April, 2025)
<https://www.ama-assn.org/system/files/prp-health-care-spending-outpaced-gdp.pdf>.
- Hudomiet, Péter, Michael D. Hurd, and Susann Rohwedder. "The Relationship Between Lifetime Out-of-pocket Medical Expenditures, Dementia, and Socioeconomic Status in the U.S." *The Journal of the Economics of Ageing* 14 (November 30, 2018): 100181.
<https://doi.org/10.1016/j.jeoa.2018.11.006>.
- Ingraham, Bailey C, Douglas Barthold, Paul Fishman, and Norma B Coe. "Caregiving for Dementia: Trends Pre-post Onset and Predictive Factors of Family Caregiving (2002–2018)." *Health Affairs Scholar* 2, no. 3 (February 16, 2024).
<https://doi.org/10.1093/haschl/qxae020>.
- Kumar, Raj G., Evan Bollens Lund, Katherine A. Ornstein, Jing Li, Kenneth E. Covinsky, and Amy S. Kelley. "Health Care Utilization and Costs in the Years Preceding Dementia Identification." *Alzheimer's & Dementia* 19, no. 12 (September 18, 2023): 5852–59.
<https://doi.org/10.1002/alz.13476>.
- Lastuka, Amy, Elye Bliss, Michael R. Breshock, Vincent C. Iannucci, William Sogge, Kayla V. Taylor, Paola Pedroza, and Joseph L. Dieleman. "Societal Costs of Dementia: 204

- Countries, 2000–2019.” *Journal of Alzheimer S Disease* 101, no. 1 (August 12, 2024): 277–92. <https://doi.org/10.3233/jad-240163>.
- Maresova, Petra, Jan Hruska, Blanka Klimova, Sabina Barakovic, and Ondrej Krejcar. “Activities of Daily Living and Associated Costs in the Most Widespread Neurodegenerative Diseases: A Systematic Review” *Clinical Interventions in Aging* Volume 15 (October 1, 2020): 1841–62. <https://doi.org/10.2147/cia.s264688>.
- National Institute on Aging. “Alzheimer’s Disease Fact Sheet,” (April 5, 2023). <https://www.nia.nih.gov/health/alzheimers-and-dementia/alzheimers-disease-fact-sheet>.
- National Institute of Neurological Disorders and Stroke. “Progress in Alzheimer’s Disease and Related Dementias (AD/ADRD) Research,” (November 29, 2024). <https://www.ninds.nih.gov/news-events/directors-messages/all-directors-messages/progress-alzheimers-disease-and-related-dementias-adard-research>.
- Park, Kisoo, Jumin Park, Young Dae Kwon, Yoonjeong Kang, and Jin-Won Noh. “Public Satisfaction With the Healthcare System Performance in South Korea: Universal Healthcare System.” *Health Policy* 120, no. 6 (January 22, 2016): 621–29. <https://doi.org/10.1016/j.healthpol.2016.01.017>.
- Shami, Elham, Jafar Sadegh Tabrizi, and Shirin Nosratnejad. “The Effect of Health Insurance on the Utilization of Health Services: A Systematic Review and Meta-Analysis.” *Galen Medical Journal* 8 (October 14, 2019): 1411. <https://doi.org/10.31661/gmj.v8i0.1411>.
- Skaria, Anita Pothan. “The Economic and Societal Burden of Alzheimer Disease: Managed Care Considerations.” *The American Journal of Managed Care* 28, no. Suppl 10 (September 1, 2022): S188–96. <https://doi.org/10.37765/ajmc.2022.89236>.
- Tay, Lyn Xuan, Siew Chin Ong, Lynn Jia Tay, Trecia Ng, and Thaigarajan Parumasivam. “Economic Burden of Alzheimer’s Disease: A Systematic Review.” *Value in Health Regional Issues* 40 (November 14, 2023): 1–12. <https://doi.org/10.1016/j.vhri.2023.09.008>.

The Impact of Childhood Trauma on the Development of Gambling Behaviors in Young Adults By Imad Chaudhary

Abstract

This article explores the relationship between gambling behavior in young adults and childhood trauma. Using peer-reviewed studies of trauma, this article explains how the different types of childhood maltreatment (i.e, emotional, physical, sexual abuse, and neglect) shapes adolescent and young adult behavior through disruptive neurodevelopment and emotional regulation. The article identifies three significant effects of childhood trauma: childhood trauma increases the chance of poor decision-making, which can lead to gambling; childhood trauma increases impulsive or compulsive behavior, which can facilitate gambling behavior; childhood trauma creates emotional vulnerabilities and other effects which can facilitate poor gambling behavior. The effects of childhood trauma on gambling behavior are further mediated by development of post-traumatic stress disorder (PTSD), dissociation, and family transmission patterns. Ultimately, this article elaborates on the behaviors of trauma-exposed youth and the need for more training of medical professionals for early screening and trauma-informed prevention methods regarding gambling addiction.

Introduction

The emerging epidemic of problem gambling among young adults is likely attributable, in part, to its psychological underpinnings, which are rarely, if ever, part of the conversation. One of the most robust predictors across virtually all addiction research is childhood trauma (witnessing and/or experiencing abuse, neglect, and household dysfunction) (Imperatori and Innamorati #44) Compelling evidence reveals that traumatic events that occurred during critical periods of brain development can alter brain functioning and produce maladaptive responses (Imperatori and Innamorati 30). In this article, we will assess how trauma early in life may affect gambling behaviors in young adulthood across three areas of study: neurobiological dysregulation; emotional regulation; and multigenerational risk. The aim of this article is to examine the manner in which early trauma leads to gambling problems in young adulthood by focusing on three interrelated lines of research: neurobiological dysregulation, emotional regulation difficulties, and multigenerational risk.

Methods

A literature search was conducted to examine research investigating the relationship between early childhood trauma and adolescent gambling behaviors. All research cited and discussed here are from peer-reviewed research articles that are remediated through the data from the previously reviewed research articles. Some search terms I used include “childhood trauma”, “childhood gambling addiction” and “addictions in youth”. All the research mentioned and described here is drawn from studies that have appeared in peer-reviewed papers, fMRI work in addition to clinical psych analyses and large-scale behavioral surveys. Together, these papers

show how trauma forms neurological, emotional and social pathways that render young adults more susceptible to gambling addiction.

Discussion

Recent studies have shown that the prefrontal cortex and amygdala -- regions of the brain related to decision-making and emotional control - can be damaged by trauma, interfering with impulse control and long-term planning (Lane 5). More specifically, a significant subset of young adults who were victimized as children are neurologically and psychologically wired for short-term emotional relief through gambling and other similar avenues. Furthermore, trauma and adaptive strategies might have been passed down intergenerationally and reinforced as many children learn from their guardians in unstable or neglecting family environments.

Psychological and Neurodevelopmental Pathways Link Childhood Trauma to Gambling

Childhood trauma disrupts normal brain development, particularly among areas that manage decision-making, emotions and impulses. Research utilizing neuro-imaging reveals that youths exposed to trauma undergo alterations in the prefrontal cortex, insula and amygdala — areas of the brain that help regulate impulse control and assess longer term consequences. For instance, Cai and colleagues (2023) demonstrated that emotional and physical neglect critically disrupts these brain regions leading to risk taking and reduction in executive function (Cai and Li). These neurobiological changes may explain why gambling, with its immediate rewards and excitement seeking, is particularly appealing to young adults who have a history of trauma. Individuals exposed to trauma early in childhood have brains that are hard-wired to be hyperreactive to immediate rewards and underreactive to consequences, a trait that makes these individuals more susceptible to using gambling as both a mood enhancer and an escape from pain. In other words, childhood trauma sets the stage for risky, reward-seeking behavior in early adulthood.

Aside from the damage to the brain, trauma survivors are also plagued by major emotional regulation disruptions; which in turn make survivors even more susceptible to a gambling addiction. Research has found that many young people who have been treated abusively in their teens — particularly with emotional abuse or neglect — also struggle to identify their feelings, and are burdened by anxiety and mood fluctuations. Imperatori and colleagues (2015) showed that a trauma-related coping strategy known as dissociation, in which individuals “detach” from discomforting emotions, is a powerful predictor of gambling severity (Imperatori and Innamorati 5). Other studies, such as that of Sharma and Sacco (2015), have also found that trauma-exposed individuals gamble to escape or self-stimulate for means of coping with depressive symptoms, stress, or loneliness (Sharma and Sacco). Because many of these individuals never learned how to cope with their emotions properly in childhood, gambling acts as a replacement for regulation. The more emotionally out of control or numb these individuals become, the more they are lured into chasing the rush, distraction or “control” that seems possible in gambling — and lose their ability to stop compulsive patterns over time.

Trauma and gambling, finally, are often connected generationally. Time and again, we hear that people who were traumatized during childhood not only gamble more as adults, but have emerged from family systems where emotional neglect, abuse or addiction was normalized. Felsher and colleagues (2009) find that problem gamblers were more likely to have experienced all types of childhood maltreatment than non-gamblers, and Lane and colleagues (2016) results supports these findings by cautioning that individuals exposed to trauma early in life could also transmit neglectful and abusive behaviors to their own children (Lane) (Felsher). This pattern is perpetuated through the generations due to lack of emotional modelling, unprocessed post-traumatic stress disorder and learned behaviour patterns either to avoid or engage without thought. In some cases, trauma survivors may not gamble to relieve present stress — but because the silence, boredom or stillness of low-stress times can bring back past traumas. Horak and colleagues (2021) study revealed that among people recently exposed to low levels of life stressors, increasing trauma history showed the strongest association with gambling, showing that the behavioral risk is less rooted in current experiences and more entrenched in unresolved old suffering (Horak 10).

Limitations

Although the present investigation has begun to positively link childhood trauma and gambling in young adults, there are numerous limitations to consider. While many studies have begun to interrogate this relationship, most are correlational and not causal – meaning they indicate association, but not evidence that suggesting trauma causes gambling behavior. Many studies also rely on participants to self-report trauma history, which could contribute to underreporting or conscious or subconscious inaccurate reporting. There are also no longitudinal studies assessing how childhood trauma relates to problematic gambling behavior development over a long period of time, limiting what we know about how trauma affects progression over time. Furthermore, most samples are based on monocultural and clinical samples, making it difficult to generalize outside of a Western clinical sample to all possible cultural contexts or socioeconomic statuses.

Future Directions

Future research should utilize longitudinal paradigms and cross-cultural participants in studies on trauma-exposed individuals into adulthood to understand when, how, and why, tendencies towards gambling behavior may emerge. Neuroimaging and genetic studies could also serve as a means to identify biological mechanisms underlying this behavior, while qualitative studies would provide richer context of the emotional and social experiences of those who have experienced trauma. Finally, coordinating mental health providers with the educational and policy partner in designing trauma-informed prevention programs or early strategies that address the antecedent conditions leading towards trauma, but also provide intervention areas and outcomes for behavioral responses to said trauma would be greatly beneficial.

Conclusion

In conclusion, the evidence shows time and again that childhood trauma influences the wiring of the brain, emotional regulation and behaviors in young adults. Neurobiological studies have indicated that trauma reshapes brain circuitry, adding impulsivity and stripping away decision-making controls. Mental health evidence suggests that trauma survivors gamble to cope with emotional distress for the purpose of distraction, self-soothing or managing psychological states. The biographical and social basis of risk involves an accumulation over generations of maladaptive coping patterns in emotional chaos. For this reason, there is a need for trauma-informed prevention and early intervention approaches, particularly among young adults, in which normative patterns of gambling sometimes do not attract attention. Detecting trauma early, easier access to mental health supports, and education on healthy content emotional regulation may mitigate gambling-based disorders. If gambling is framed as a symptom of trauma, rather than a moral failing or minor diagnostic category, practitioners may be more equipped to disrupt and stop the cycle before it progresses.

Works Cited

- Cai, Jianjian, and Jia Li. "Long-term effects of childhood trauma subtypes on adult brain function." *Brain and Behavior*.
- Felsher, Jennifer. "Young Adults with Gambling Problems: The Impact of Childhood Maltreatment."
- Horak, Natascha. "Gambling Disorder and Childhood Trauma: A Complex Association." *Journal of Gambling Studies*.
- Imperator, Claudio, and Marco Innamorati. "The Association among Childhood Trauma, Pathological Dissociation and Gambling Severity in Casino Gamblers." *Clinical Psychology and Psychotherapy*.
- Lane, Wendy. "Child maltreatment and problem gambling: A systematic review." *Elsevier*, 2015.
- Sharma, Anjalee, and Paul Sacco. "Adverse Childhood Experiences and Gambling: Results From a National Survey." *Journal of Social Work Practice in the Addictions*.

Cost-Benefit Analysis of Climate Change Adaptation Strategies in Portugal, Mexico, and South Korea By Ramilya Makhmudova

Abstract

One major global problem that has brought serious issues to the environment, especially in recent years, has been climate change. Climate change, or a long-term shift in global temperature and weather patterns, has negatively affected nearly all sectors of society, as well as destabilized countless cities and neighborhoods vulnerable to its effects. Due to the rising sea levels in recent years due to climate change, coastal cities in numerous regions have begun implementing infrastructure to adapt to the new environment. This paper reviews, analyzes, and compares the benefits of different climate change adaptation strategies and their associated costs on the coasts of three continents in countries such as Portugal, Mexico, and South Korea. This review intends to 1) account for the costs of damage due to the consequences of climate change, 2) analyze the effectiveness of various climate change adaptation strategies, and 3) present the most optimal strategies for each of the study areas. In this literature review, the work is organized as follows. Section 2 presents an overview and gives a description of the study areas. Section 3 analyzes findings and data. Section 4 discusses the findings and gives recommendations for each of the three study areas. Section 5 lists acknowledgements.

1. Introduction

The increasing level of greenhouse gas emissions due to a combination of both human and natural activities has immensely contributed to climate change, making it the world's largest environmental issue at hand. Coastal areas, or areas located near larger bodies of water, have become especially at risk of natural disasters such as flooding due to rising sea levels. In fact, sea levels have begun increasing ever since 1880, and are only continuing to rise (Lindsey, 2023). In an effort to adapt to these changes, coastal cities have begun to implement infrastructure, as well as utilize other adaptation strategies, such as accommodation and protection, to minimize the risk of flooding. However, each strategy comes with its own associated cost and efficiency; therefore, this research analyzes a multitude of adaptation strategies, and, in comparing the costs and benefits of each, suggests the most effective and least costly options.

The existing literature on the various climate change adaptation strategies discusses a vast range of important topics within the research field. Previous studies have emphasized different approaches to adaptation; some have shed light on the importance of culturally appropriate adaptation, local awareness, and the need to cater adaptation strategies to each individual region (Bonnet and Birchall, 2020, Escudero and Mendoza, 2021, Sinay and Carter, 2020). Studies also showcase the role of the government, calling attention to decentralized decision-making for more effective adaptation. Among climate change adaptation literature, adaptation strategies have been categorized into different types, such as hard and soft adaptations, and analyzed to better understand their costs and benefits, highlighting the necessity to economically evaluate each strategy (Cardona et al., 2020, Golub and Golub, 2016, Woodruff et al., 2018). In coastal regions

with city layouts, studies push the importance of urban planning to best construct fitting adaptation strategies (Lee, 2014, Abdulrahman et al., 2024, Roukounis et al., 2024). Studies have even begun looking into the future for hybrid solutions such as combinations of ecological and engineering approaches to combat rising sea levels and flooding (Hübner, 2025, Roukounis et al., 2024, and Abdulrahman et al., 2024). Along with the implementation of different strategies, researchers have looked into the equity and justice associated with the areas at risk, concluding that justice is a vital to implement these strategies (Okomoto and Doyon, 2024, Babapoorkamani et al., 2025, Transformational Adaptation Review, 2021). Lastly, research on the emerging use of AI in infrastructure planning is beginning to surface in the field, suggesting its use for a more efficient adaptation process. Costa et al., 2024 proposes the use of reinforcement learning for infrastructure and urban planning.

2. Study Areas

This paper studies the climate change adaptation strategies present in three coastal regions at risk: Tabasco, Mexico, bordering the Gulf of Mexico Furadouro Beach, Ovar Portugal, bordering the Atlantic Ocean, and Mokpo, Korea, bordering the Yellow Sea. All of the studied coastal areas are at significant risk of flooding and sea level rise and have been majorly affected by the consequences of climate change. Furadouro Beach in Over, Portugal, has seen an average retreat of 8 meters from 1984 and 1990, an increase from the 2.8 meter retreat between 1954 and 1980, suggesting a shoreline loss (Pereira and Coelho, 2013). The area is vulnerable to overtopping and coastal erosion, which can lead to a continued loss of assets (Cruz et al., 2015). The sea level rise rate has also more than doubled from 1.5mm/year between 1882 and 1990s, to around 4.1mm/year between 2005 and 2015 (Cardona et al., 2020). 40% of the city of Mokpo, Korea is predicted to be inundated by 2100 due to rising sea levels (Lee, 2014). Having one of the most diverse populations of mangroves in the world, in addition to diverse flora and fauna on beaches and in dunes, coastal regions of Mexico have become vulnerable to the negative effects of natural disasters such as cyclones and storms. These natural disasters have exacerbated coastal erosion and flooding, decreasing biodiversity and hurting animal habitat (Escudero and Mendoza, 2021). Of all states in Mexico, Tabasco has received the most precipitation, increasing chances of flooding and putting the region at risk of erosion and overtopping, averaging 2095 mm/year between 1971-2000. The coastal region is also vulnerable to storms and hurricanes coming from both the Pacific Ocean and the Gulf of Mexico (Haer, 2017). The causes of these issues are strongly linked to anthropogenic actions, exacerbating and contributing to rising sea levels that are projected to accelerate in the coming years. Infrastructure built by humans for the purpose of protecting waters from entering buildings, for instance, has turned soil more prone to erosion, made soil more impermeable, and harmed vegetation (Cardona et al., 2020).

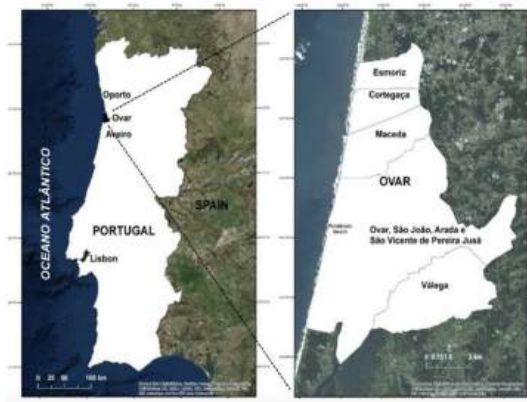


Figure 1. Study Area 1. Coastal city, Ovar, in Portugal bordering the Atlantic Ocean (Cardona et al., 2020).



Figure 2. Study Area 2. Coastal State of Mexico, Tabasco, bordering the Gulf of Mexico (Haer et al., 2017).



Figure 3. Study Area 3. Coastal city, Mokpo, located on the southwestern tip of the Korean Peninsula. The city borders the Yellow Sea to the West.

3. Data Analysis

This study sorts adaptation strategies into three categories 1) protect 2) accommodate 3) retreat. Strategies categorized under “protect” are associated with implementing hard or soft engineering structures. “Accommodate” refers to the non-structural, biophysical, and institutional responses to the potential risk. Relocation of people and infrastructure, as well as other assets is categorized as “retreat.” Some literature, such as that by Lee (2014), also include “attack” as a category that refers to the expansion of coastal areas seaward via infrastructure, such as piers and bridges, in an effort to adapt to changing sea level heights.

Table 1. Monthly defense (protection) costs in 2014 in Study Area 1, Ovar, Portugal (Cardona, 2015).

Defense structures	Maintenance costs (€/m)
Groin	2709,30
Seawall	2438,37
Sand dyke	866,56

Cardona et al. (2015) accounts for three protection structures: Groin, Seawall, and Sand dyke. The monthly maintenance costs are presented in Table 1. The research also accounts for upkeep, or maintenance cost over time, using standardized unit costs and an inflation rate of 3% as a discount rate for 50 years. Accommodation costs were calculated as 30% of a construction cost for a building, if building dimensions and numbers remained constant. Buildings categorized as at high or very high risk were considered, with an inflation of 3% and accommodation over 25 years. Retreat costs were calculated using previous building construction costs and the building demolition costs, considering only high or very high risk buildings. (Cardona et al., 2015).

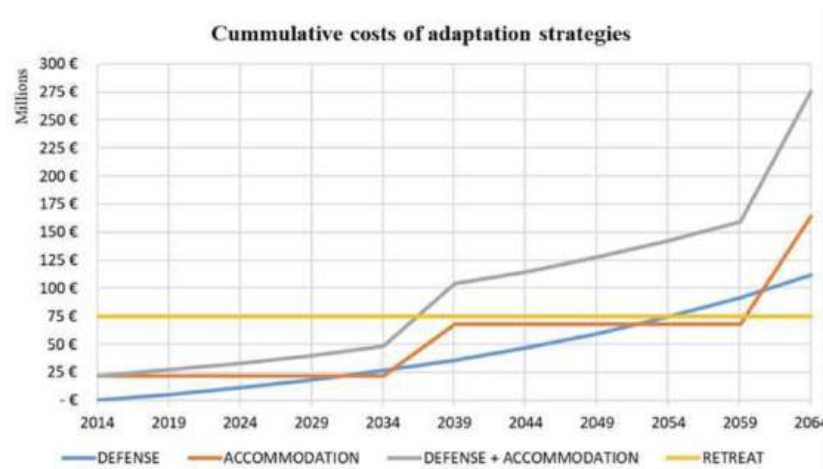


Figure 4. Different adaptation strategies, including combined strategies, graphed with direct associated costs (Cardona et al., 2015).

Table 2. Indirect costs for adaptation strategies (Cardona et al., 2015).

Adaptative Strategies Indirect Costs		
Defense	Accommodation	Retreat
• Interferences in the coastal dynamics; • The seawall increases local erosion, as the sea, when striking these structures, causes return currents that remove the sediments; • Visual impact and loss of landscape value; • Downdrift site erosion.	• It doesn't reduce the danger, putting the inhabitants at risk; • If the defense infrastructures are not maintained, the danger of overtopping and flooding will increase.	• Coastline retreat, with loss of land; • Change in land use; • Habitat change; • Community hesitation in relocation, due to local traditions and landscapes, as this strategy implies changes in communities and lifestyles.

Table 3. Benefits of adaptation strategies (Cardona et al. 2015).

Adaptative Strategies Benefits		
Def.	Accomm.	Ret.
• Coastal population defense from the sea; • Maintains the coastline in the areas with seawall defense, preventing its retreat; • The groins retain sediments updrift, carried by the coastal drift increasing the width of the beach in these areas; • Decreased frequency of coastal flooding; • Habitat gain (groins); • Possibility of the coastal community continuing to live in the same places.	• Reduce the risk to population and assets, as the community adapts and increases the capacity to deal with the impacts • Able to continue to live in the same places; • Not change the coastal dynamics.	• This strategy is the one that most minimizes the risk, since the buildings at risk are relocated to areas without sea hazards; • Not built public spaces gains, with the dual function of being barriers to the sea actions and, at the same time, used as recreation and leisure zones; • Permeabilization of the soil increase reducing the dangerousness of the coastal flooding; • Renaturation also increases sediment retention and, therefore, reduces coastal erosion; • Habitats previously occupied by the buildings reappear; • Reduce the costs of the damages done by coastal overtopping and flooding; • Restoration of natural environmental heritage; • Increase in landscape and environmental value; • Education and science potential; • Does not change the coastal dynamics;

The graph illustrated in Figure 4 presents the costs of each adaptation strategy. The price for the “retreat” strategy remains constant throughout the years, standing at a direct cost of €75M after 50 years. The “accommodate” and “defend (protect) + accommodate,” are the highest adaptation strategies in terms of cost (€164M and €267M, respectively). The “defend” adaptation strategy is projected to cost €112M in the next 50 years (Cardona et al., 2015). Table 3 presents indirect costs and negative consequences, not accounted for in calculations, of adopting any of the three adaptation strategies. The cross-over times of each adaptation strategy with “retreat” indicate when it will become cheaper to relocate than adopt other strategies. These cross-overs showcase that the prices of seemingly cheaper alternatives can accelerate over time, jumping past the retreat strategy and eventually costing more due to maintenance or other indirect costs.

Coastal regions in Mexico have seen large-scale damage as a result of flooding caused by sea level rise. Tabasco, a coastal state in Mexico bordering the Gulf of Mexico has been affected by two types of flooding – coastal and river flooding – and has been vulnerable to the economic damage associated with the phenomenon (Haer et al., 2017). The current damage of \$0.53B due to coastal flooding is projected to nearly quadruple, reaching an Expected Annual Damage (EAD) of \$4.12B if not adaptation strategies are adopted. The current damage of \$1.79B due to river flooding, on the other hand, is expected to rise by nearly 10 fold, to \$10.6B (Haer et al., 2017).

Table 4. Direct and indirect damage to coastal lands in Tabasco as a result of flooding, hurricanes, and droughts, quantified in millions USD between 2007 and 2011. Reported by the United Nations Economic Commission for Latin America and the Caribbean (CEPAL, 2011).

Year	Direct damage	Indirect damage	Total damage	Cumulative damage %
Million USD				
2007	1509	1192	2701	56
2008	270	177	447	9
2009	143	66	209	4
2010	441	186	626	13
2011	647	226	873	18
Total	3009	1847	4857	100

Table 5. Hectares of Mangroves, average mangrove restoration costs, present value, present value per hectare, cost-benefit ratio, and annual expected people protected by mangrove restoration efforts in Mexico at a country scale, calculated at a 4% discount rate for 30 years (Menéndez et al., 2022).

Hectares of Mangroves	Median Mangrove Restoration Cost (USD/hectare)	Present Value (USD million)	Present Value per Hectare (USD)	Benefit to cost ratio (B/Cr)	Annual Expected People Protected
940, 170	4, 538	6, 260. 61	6,659	1.47	110, 243

Mangrove restoration falls under the “protect” adaptation strategy category, as it blocks water from entering coasts. Table 5 quantifies the benefits, values, projected people protected, and mangrove restoration costs per hectare for Mexico. Table 5 provides a cost-benefit ratio of 1.47, implying that ecological-based protection, or protection via mangrove restoration, is an efficient adaptation strategy that successfully minimizes potential risk in Mexico on a country scale. In addition, Menéndez et al. (2022) predicts that over 110,000 people on coasts are

protected from sea level rise, as well as key assets such as Villahermosa City and small and medium-sized coastal recreational and fishing communities in Tabasco state (Rogers et al., 2022).

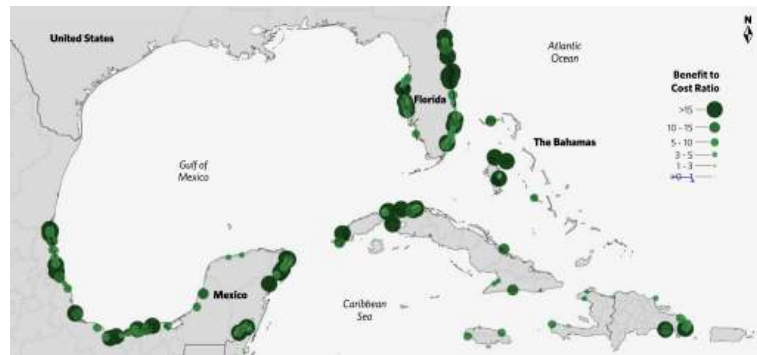


Figure 5. Benefit to cost ratios of mangrove restoration for 5-km coastal study units, at a 4% discount rate for 30 years in areas bordering the Gulf of Mexico and the Atlantic Ocean, assuming a median restoration cost of \$4,538 per hectare for Mexico (Menéndez et al., 2022).

As presented in Figure 5, the coastline of Tabasco state, Mexico maintains a consistent, high BCR, highlighting mangrove restoration, or “protection” as an efficient and beneficial adaptation strategy.

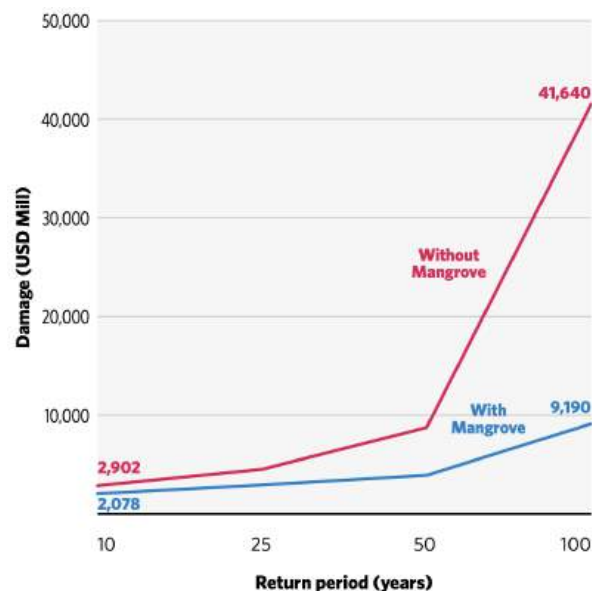


Figure 6. Coast damage with (blue line) and without mangroves (red line) per return period (years) due to flooding in Mexico (Menéndez et al., 2022).

Implementing “protect” infrastructure, such as mangroves or other coastal ecosystems, benefits that counteract the negative effects of sea level rise. Mangroves, an example of ecosystem-based solution, can serve as natural barriers to rising sea levels, large waves, or storm

surges, minimizing coastal flooding or overtopping. Similarly, there are other benefits that can contribute to and better the quality of life of local communities, such as carbon sequestration and food provisioning (Menéndez et al., 2022).

The costs for adaptation strategies under “accommodate” and “retreat” in coastal regions of Mexico have not been explicitly explored or published in existing literature. However, Haer et al. (2017) claims that improved forecast systems, warnings, and evacuation planning for coastal communities can better prepare the public for potential hazards, minimizing the resulting casualties. “Retreat” or migration costs can be determined by multiplying 3 by local GDP per capita per migrant, or 2.8 by local GDP per capita for each resident of the area (Hallegatte et al., 2013), which yields an estimate of the cost of the assets on lost land, as well as a deconstruction cost of 7% (Diaz, 2016). For Mexico, at a country-scale, the “retreat” cost would equal roughly \$39,642, using Mexico’s most recent GDP-pc (World Bank Open Data). At a state-scale, the “retreat” cost would equal approximately \$65,310, using Tabasco’s most recent Gross State Product (OECD). There are no explicit benefits or numerical B/C ratios for “retreat” or migration strategies. However, taking into account the effectivity of migration strategies in other areas, such as Portugal, and comparing the “retreat” option with no “retreat” option, or the current and previous damage costs presented in Table 4, “retreat” options are a notable way to lower SLR (Sea-Level Rise) costs and combat future loss of assets (Lincke and Hinkel, 2021).



Figure 7. Likely inundation zones in and near Tabasco, Mexico. Zones shaded red represent low-lying risk areas below 5ft of water with an unobstructed path to the ocean (Climate Central, 2024).

Figure 7 highlights low-lying, flood-prone areas near Tabasco, Mexico; as indicated by the map, with data based on global datasets for elevation, tides, and sea level rise projections, Tabasco is an area prone to flooding, with land below 5ft of water. Communities in regions at risk may majorly benefit from relocating from high risk to lower risk areas, or areas not shaded in red, thereby protecting themselves from future natural hazards. “Retreat” may serve as a beneficial way to decrease loss of assets due to flooding, and the associated costs of SLR (Lincke and Hinkel, 2021).

Because Korea’s shores are surrounded by water on three sides the public has begun taking advantage of this open access to water bodies. Korea has begun building infrastructure

and setting up construction projects for recreation, dwelling and power generation. However, due to natural phenomena such as flood events and increasing wave heights, shorelines and coastal communities have become more vulnerable to the damage of the newly-built infrastructure (Lee et al., 2014). It is estimated that the damage cost can reach up to \$20.6B in 2060, at a country level (Lee et al., 2017). In addition, precipitation during the summer over the past 30 years has increased by an average of 0.1mm per year (Korea Meteorological Administration, 2020).

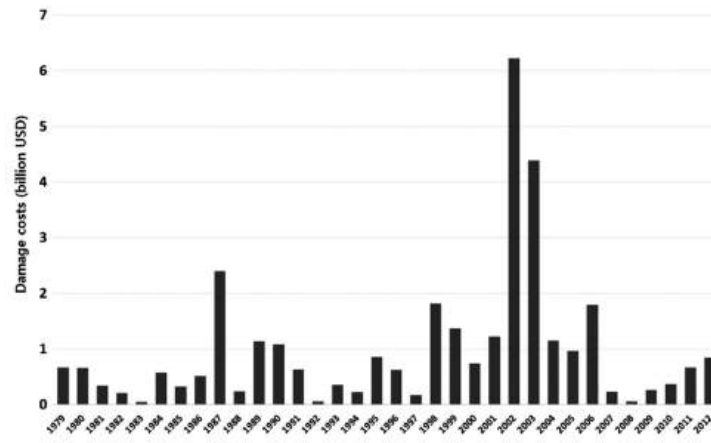


Figure 8. Damage costs for Korea (billion USD) between 1979 and 2012 (Lee et al., 2017).

Figure 8 presents the damage costs for the loss of assets between the years 1979 and 2012. Damage costs peaked in the year 2002, reaching over \$6B. In the fall of 2002, the Korean peninsula was hit by a large typhoon, the most severe in Korean history (ADRC).

Table 6. Different adaptation strategies in Mokpo, Korea, including “protection,” “accommodation,” “retreat,” and “attack,” as well as their features, methods of use, advantages (benefits) and diagrams (Lee et al., 2014).

Strategies	Features	Methods	Site Advantages	Diagrams
Protection (Defense)	Hard Dikes Levees Sea Walls Groins	Build physical barriers to block water	For hard-to-move facilities and infrastructure on flat ground	
	Soft Mangroves Wetlands Sand dunes Tidal Flats	Create buffer with vegetation or landforms	For maintaining shoreline at sites with existing coastal forest or sand dunes	
Accommodation	Raising level Desalination Drainage Alarm system	Upgrade functions while maintaining sites without location	For redevelopment projects or facilities at high ground nearby	
Retreat	Relocation Abandonment	Relocate facilities to low-risk uplands	For residential and public facilities at sites with low-risk uplands nearby	
Attack	Land Reclamation Piers Ports Harbors	Extend facilities towards water	For facilities requiring direct access to water	

Table 6 presents 17 different adaptation strategies within four adaptation categories. Under “protection” are two sub categories, soft and hard. Soft protection strategies include ecosystem-based solutions such as mangroves, wetlands, sand dunes, and tidal flats. Hard “protection” infrastructure refers to anthropogenic engineering (dike levees, sea walls, groins). Under “accommodation” are adaptation efforts as well as efforts to better inform the public in case of emergencies. The "retreat" category involved two processes, relocation and abandonment, both of which refer to the movement away from the shores. Lee et al. (2014) also considers “attack,” referring to strategies that expand seaward to adapt to increasing water levels.

Green spaces are another adaptation strategy that fall under soft engineering solutions in the “protect” category. Green spaces act as natural buffers and filters, offering benefits such as protection of coasts from storm surges and sea level rise due to climate change. Other co-benefits can include fostering greater biodiversity and protecting habitat. Park et al. (2023) estimates that green space construction costs for every 625 m² amount to \$41,500. Seawalls, on the other hand, categorized as hard “protection” strategies, were estimated to cost about \$550 per linear meter of height and length. Thus, with \$41,500, or the same amount of money, a seawall of 75m in length and 1m in height could be built, making seawalls a more costly adaptation option.

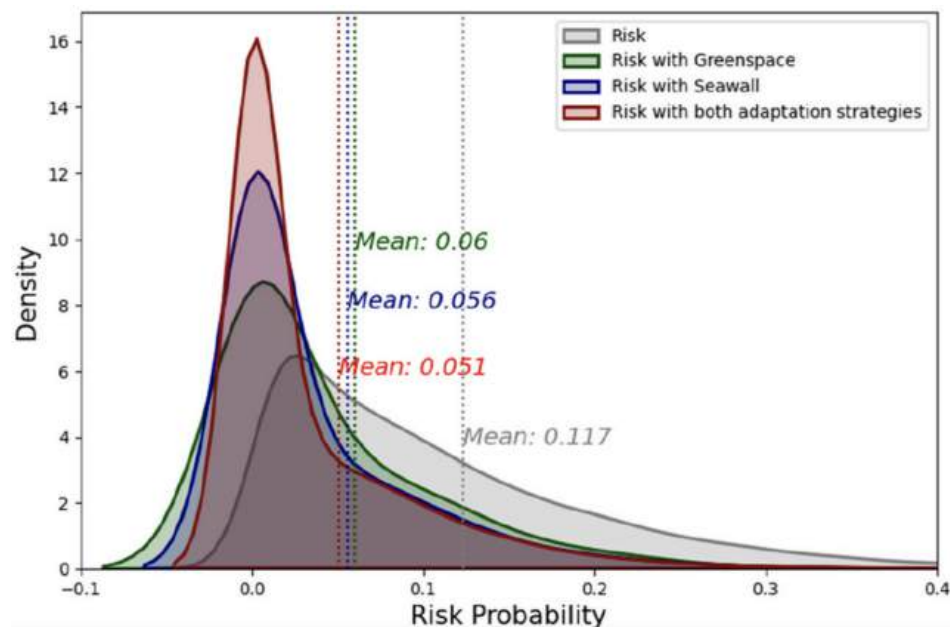


Figure 9. Different coastal flooding risk probabilities for different adaptation strategies.

As presented in Figure 9, implementing green spaces, or soft infrastructure is more effective at reducing risk than seawalls, or hard infrastructure (Park et al., 2023).

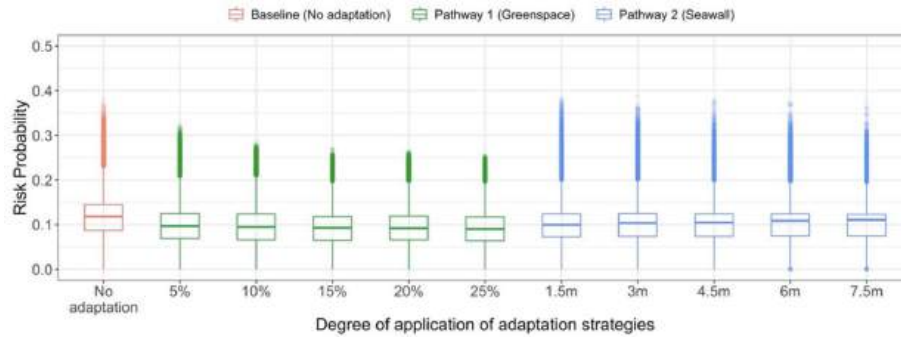


Figure 10. Flooding risk probability based on the degree of application of each adaptation strategy (green space, seawall, no adaptation) (Park et al., 2023).

Figure 10 refers to the degree of application and the associated risk probabilities. Green spaces are generally more efficient at reducing risk than seawalls, especially when adopted at a higher degree.

Among the current literature about climate change adaptation strategies, particularly under categories such as "accommodation" and "retreat," explicit, quantitative costs for implementation of each are yet to be published. However, applying the same relocation rule used for Mexico's "retreat" costs, (Diaz, 2016, Hallegatte et al., 2013), Korea's "retreat" costs would equal to roughly \$92,738, on a national scale, using the its most GDP-pc (World Bank Open Data). Individual GDP-pc for Mokpo, Korea are not reported separately, and thus exact "retreat" costs cannot be derived from it, but other coastal cities in Korea, such as Busan and Ulsan, have approximate "retreat" costs of \$88,510 and \$217,031 (World Bank Open Data, KOSIS, 2022).

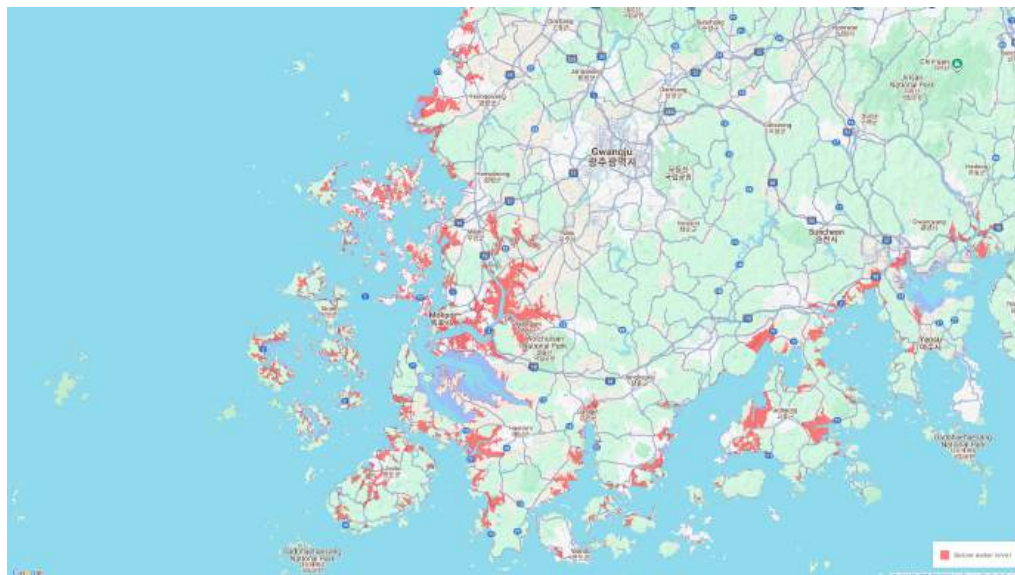


Figure 10. Likely inundation zones in and near Mokpo, Korea. Zones shaded red represent low-lying risk areas below 5ft of water with an unobstructed path to the ocean (Climate Central, 2024).

Figure 10 presents flood-prone coastal areas below 5ft near Mokpo, Korea (shaded red). The data represents areas where local assets are at risk of loss due to climate change hazards.

Thus, in implementing the “retreat” adaptation strategy – which involves relocation from flood-prone coastal areas to less hazardous regions – can serve as an efficient way to minimize potential losses of assets and the associated SLR costs.

4. Discussion and Recommendations

Although it is evident that the coasts of each of the three study areas have faced similar issues – overtopping, erosion, precipitation, storm surges – due to climate change, each of the regions is also suffering from the consequences at different degrees. That is to say that every region would benefit from a set of next steps or solutions that caters specifically to the most prevalent issues at hand in that community.

Figure 1 detailing the costs of each adaptation strategy from Study Area 1, Ovar, Portugal suggests that the costs for each of the adaptation strategies is projected to accelerate in the coming years. Considering the projected increases in costs over time, the most optimal solution to the sea level rises in Ovar would be a combination of multiple strategies, or the practice of swapping strategies over time. Ultimately, any strategy (accommodation, protection, or a mix of the two) would be cost-efficient if swapped with retreat after the costs of the two overlap in the coming years. Otherwise, “retreat” alone is a strategy that guarantees a consistent price over time, and is the most efficient at minimizing risks of coastal flooding, as presented in Table 3.

Table 5 quantifies the present values of the costs and benefits of mangrove restoration in Study Area 2, Mexico on a country-scale, a “protection” strategy, and gives a benefit-cost ratio of 1.47, indicating a worthwhile project, likely to be successful in battling flooding by protecting communities at risk. More specifically Figure 5 details the BCR approximations, with a 4% discount rate for 30 years summarized at 5-km coastal study units, for regions bordering the Gulf of Mexico and Atlantic Ocean, including Tabasco state, an area with a consistently high BCR (Menéndez et al., 2022). In addition, Mexico’s mangroves population, the largest and most diverse in the world, have been at risk due climate change and anthropogenic actions; with the implementation of mangrove restoration efforts, mangroves will remain protected from risks of extinction or other potential harm. The “retreat” option, though more costly (~\$65, 310, calculated by 3xGDP-pc), can decrease SLR costs and losses of assets, and efficiently spare coastal communities the burden of facing flooding and its consequences entirely (Lincke and Hinkel, 2021). Thus, adopting adaptation strategies such as “protect” and “retreat” would both shelter communities and benefit the surrounding environment, as well as minimize any future losses.

For Study area 2: Korea, being lower in cost and more efficient, green spaces are the optimal solution to protect communities from increasing wave heights and floods. However, the coastlines in South Korea could benefit from a mix of both green spaces and seawalls, as well. Though higher in price, seawalls are still an efficient structural approach to protect against coastal disasters, especially in lower-risk areas where they can be used in smaller quantities rather than large-scale buffers spanning along large distances (Park et al., 2023). In addition, the “retreat” option is a highly efficient way to reduce damage costs and minimize risk for

communities in Mokpo. Similarly to Mexico and Portugal, Korea's coastal regions could consider migrating to areas less prone to flooding and overtopping. In doing so, communities would remain more protected and continuous losses of assets and damage costs would be reduced.

5. Conclusion

This review has performed a cost-benefit analysis for each of the identified study areas, Portugal, Mexico, and South Korea, also detailed in Section 2. The study has concluded that the optimal way to address the effects of climate change in each region is to examine the most prevalent issues, assess costs, and compare them with the impacts, or the benefits of each strategy. It is important to reiterate that as the climatic state of each country or region experiences change, the optimal climate change adaptation strategies may change. This review also strives to contribute to the existing literature on climate change adaptation strategies, by emphasizing that adaptation comes in various forms – mixed approaches and switching approaches, for instance – that address specific issues in a certain community, and thus oftentimes cannot be solved in one particular way or with a generalized solution that works for other regions.

Though this review presents a comprehensive view of most costs and benefits of the proposed strategies, as well as general overviews of each of the studied areas, this study does not include CBAs for “accommodation” strategies for Korea and Mexico, and BCRs for each strategy, as quantitative data on these literature topics for the studied regions is scarce. These gaps, however, can potentially become future research topics that could contribute to environmental and economic research fields. In addition, this study does not suggest that CBA is the only way to assess climate adaptation solutions, and does not apply other decision criteria outside of BCR, but rather serves as a tool for comparison and insights, as it is used widely across various literature.

6. Acknowledgements

This review was supported by Ceyhun Elgin, a professor at Columbia University and mentor at the Indigo Research Program.

Works Cited

- Abdulrahman, H., Al-Ghamdi, S. G., & Al-Saadi, A. (2024). Global strategies for climate adaptation in coastal cities. [ResearchGate].
<https://www.researchgate.net/publication/386090329>
- Babapoorkamani, M., Colenbrander, D., & Simatele, D. M. (2025). Climate adaptation decision-making in African coastal regions: A systematic review. *Climate Risk Management*, 46, 100596. <https://doi.org/10.1016/j.crm.2025.100596>
- Bonnett, N., & Birchall, S. J. (2020). Coastal communities in the Circumpolar North and the need for sustainable climate adaptation approaches. *Marine policy*, 121, 104175. <https://doi.org/10.1016/j.marpol.2020.104175>
- Cardona, F. S., Ferreira, J. C., & Lopes, A. M. (2020). Cost and Benefit Analysis of Climate Change Adaptation Strategies in Coastal Areas at Risk. *Journal of Coastal Research*, 764–768. <https://www.jstor.org/stable/48748800>
- Climate Central. (2024). Coastal Risk Screening Tool: Sea level rise and coastal flooding impacts [Interactive map]. Climate Central. <https://coastal.climatecentral.org/>
- Cruz, T., Pereira, C., Coelho, C., & Roebeling, P. (2015). Desempenho de obras longitudinais aderentes no concelho de Ovar, Portugal [Performance of adherent longitudinal works in the municipality of Ovar, Portugal]. Paper presented at the VIII Congresso sobre Planeamento e Gestão das Zonas Costeiras dos Países de Expressão Portuguesa. http://www.aprh.pt/ZonasCosteiras2015/pdf/1B2_Artigo_045.pdf
- Diaz, D. B. (2016). Estimating global damages from sea level rise with the Coastal Impact and Adaptation Model (CIAM). *Climatic Change*, 137(1–2), 143–156. <https://doi.org/10.1007/s10584-016-1675-4>
- Escudero, M., & Mendoza, E. (2021). Community Perception and Adaptation to Climate Change in Coastal Areas of Mexico. *Water*, 13(18), 2483. <https://doi.org/10.3390/w13182483>
- Golub, A., & Golub, E. S. (2016). Major interventions (with benefit and cost estimations). In *Cost-Benefit Analysis of Adaptation Strategy in Bangladesh* (pp. 25–56). Copenhagen Consensus Center. <http://www.jstor.org/stable/resrep21826.6>
- Fernandes, R. (2015, October). Gestão de emergências em zonas costeiras [Conference paper]. In *Zonas Costeiras 2015*. APRH. https://www.aprh.pt/ZonasCosteiras2015/pdf/1B2_Artigo_045.pdf
- Hallegatte, S., Green, C., Nicholls, R. J., & Corfee-Morlot, J. (2013). Future flood losses in major coastal cities. *Nature Climate Change*, 3(9), 802–806. <https://doi.org/10.1038/nclimate1979>
- "Regional Economy: Gross Domestic Product, Large Regions". OECD Statistics. Retrieved 2021-01-26.
- Korea Meteorological Administration. (2020). Korea climate change assessment report 2020. Korea Meteorological Administration.
- KOSIS statistical table: Regional GDP, gross regional income and individual income

- Statistics Korea. (2024, May 28). Regional GDP, gross regional income and individual income (Annual 1985–2022). KOSIS.
https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1C86&conn_path=I2&language=en
- Lee, M., Hong, J. H., & Kim, K. Y. (2017). Estimating damage costs from natural disasters in Korea. *Natural Hazards Review*, 18(4), Article 04017016.
[https://doi.org/10.1061/\(ASCE\)NH.1527-6996.0000254](https://doi.org/10.1061/(ASCE)NH.1527-6996.0000254)
- Lincke, D., & Hinkel, J. (2021). Coastal migration due to 21st century sea-level rise. *Earth's Future*, 9, e2020EF001965. <https://doi.org/10.1029/2020EF001965>
- Lindsey, Rebecca. (2023) National Oceanic and Atmospheric Administration (NOAA). Climate change: Global sea level. Climate.gov. Retrieved August 13, 2025, from <https://www.climate.gov/news-features/understanding-climate/climate-change-global-sea-level>
- Menéndez, P., Lowrie, C., & Beck, M. W. (2022). Building Mangrove Capital: Assessing the benefit–cost ratio for mangrove restoration across the wider Caribbean (The Nature Conservancy report). The Nature Conservancy. Retrieved August 13, 2025, from https://www.nature.org/content/dam/tnc/nature/en/documents/TNC_MangroveCapital_140922.pdf
- Park, D. (n.d.). Disaster damage report by Typhoon Rusa in S. Korea. Ministry of Government Administration and Home Affairs, Republic of Korea. In ADRC counterpart report: Korea (pp. 1). Retrieved August 13, 2025, from https://www.adrc.asia/counterpart_report/korea0209/korea_0209_01.pdf
- Sinay, L., & Carter, R. W. (2020). Climate Change Adaptation Options for Coastal Communities and Local Governments. *Climate*, 8(1), 7. <https://doi.org/10.3390/cli8010007>
- Woodruff, S., BenDor, T.K. and Strong, A.L. (2018), Fighting the inevitable: infrastructure investment and coastal community adaptation to sea level rise. *Syst. Dyn. Rev.*, 34: 48-77. <https://doi.org/10.1002/sdr.1597>
- World Bank. (n.d.). World Bank Open Data. Retrieved August 24, 2025, from <https://data.worldbank.org/>

Analyzing Stock Market Behavior: Empirical Evidence From the United States

By Zachary Leeds

Abstract

This empirical paper investigates the impact of the COVID-19 pandemic on stock market volatility and returns in the United States from 2020 to 2023. This paper analyzed 6 major sectors: Aggregate Stock Market Index, Communication Services, Technology and Information, Health Care, Consumer Discretionary, and Industrials. The result of this study indicates that there was extreme volatility on March 11, 2020, when COVID-19 was first announced as a global pandemic by the World Health Organization. We can see that investors' sentiment shifted to digital and online-based companies during the pandemic period, such as Microsoft and Amazon. The findings reflect the sector-dependent nature of volatility response to the COVID-19 pandemic, suggesting the importance of diversification in portfolio management.

Keywords US Stock Market, Volatility, Return, Regression, COVID-19

Introduction

Understanding the stock market behavior during the COVID-19 period is important because the financial market (NASDAQ) correlates highly with the global aggregate economy. Volatility is a key indicator for financial theory in terms of VaR calculation, hedging strategies, portfolio allocation, and for the overall aggregate economy. Investigating market volatility among multiple sectors during a period of crisis is important to understand the vulnerability of other sectors during a period of uncertainty. This insight is important for investors to construct the necessity of having a diversified portfolio and policymakers when reducing systemic risk. This study focused on the United States, which is one of the world's largest and most influential stock markets. By focusing on the 6 main sectors of the US stock market, we created a small sample that represents the entirety of the market during the pandemic period. Stock volatility provides valuable insight into how investors react to unforeseen events, underscoring the unpredictability of the stock market. We focused on COVID-19 because the pandemic affected populations globally; lockdown declined production, collapsing demand, and caused disruption and financial uncertainty worldwide.

The purpose of this research is to investigate COVID-19's effect on market volatility and returns. By using regression analysis and identifying descriptive statistics, the results indicate that the COVID-19 pandemic had statistically significant positive effects on the volatility of companies in the consumer discretionary and technology & information sectors. While there is a statistically significant negative effect on the volatility of companies in the following sectors: communication services, healthcare, and industrials, after controlling for the effect of the aggregate stock market index (NASDAQ). The findings emphasize sector heterogeneity in response to the COVID-19 pandemic period.

This paper contributes to the growing literature on stock market behavior during the COVID-19 period by providing sector-specific empirical evidence from the United States' stock market.

Literature Review

Various methodologies have been used and employed to analyze stock market behavior, such as regression analysis (Uddin et al., 2021), GARCH (Bora & Basistha, 2020; Li et al., 2022; Wang et al., 2024), EGARCH (Rakshit & Neog, 2022; Khan et al., 2023), Event Study (Fernandez-Perez et al., 2021; Endri et al., 2021), HAR-USRV-VIX (Li et al., 2021), DCC-GARCH (Yiming et al., 2024), and Diebold and Yilmaz Methodology (Mao et al., 2024).

Many studies have focused on various countries and regions, such as the BRIC economies (Rakshit & Neog, 2022), Indonesia (Endri et al., 2021), Pakistan (Wang et al., 2024), France (Li et al., 2020), UK (Khan et al., 2024), India (Mao et al., 2024), Ghana (Insaiddoo et al., 2021), E-7 Countries (Yiming et al., 2024), China (Mao et al., 2024), the United States (Papadamou et al., 2020), and Global (Uddin et al., 2021).

Many Variables were used in the literature to analyze financial market behavior, such as Stock Market Volatility (Uddin et al., 2021), Exchange rate (Rakshit & Neog, 2022), GDP (Li et al., 2022), Stringency Index (Fernandez-Perez et al., 2021), Realized Volatility (Li et al., 2020), Gold (Yiming et al., 2024), and Crude Oil (Mao et al., 2024).

The COVID-19 Pandemic created unexpected uncertainty across global stock markets; several studies examined its effect on financial markets. For example, Uddin et al. (2021) found that factors like governance quality, health systems, productivity, and financial institutions play key roles in reducing volatility. Another example is Li et al. (2022) found that COVID-19 disrupted supply chains, crashed stock markets, and reduced GDP worldwide. Recovery was dependent on governmental policies, healthcare, and MSME support for sustainable growth. Endri et al. (2021) revealed that COVID-19 drove high volatility and falling stock prices in Indonesia, with GARCH models showing negative abnormal returns and opportunities for speculators to invest in the stock market. Wang et al. (2024) found that COVID-19 triggered short and long-term stock market volatility in Pakistan, worsened by lockdowns. Both Khan et al. (2023, 2024) pointed to heightened volatility in oil prices and Bitcoin. Similarly, Papadamou et al. (2020) linked investors' attention to the volatility increase in European markets.

The Literature Review agrees that the COVID-19 pandemic has a significant effect on Stock Market Volatility worldwide, which varies depending on countries' characteristics and sectoral exposure. The GARCH class models were most frequently used to capture these volatility patterns.

Authors	Year	Key Variable	Methodology	Key Findings	Time Period	Countries/R egions
Uddin et al.	2021	Governance & Volatility	Regression analysis	Strong governance reduced volatility during COVID-19.	2019-2021	Global

Bora & Basistha	2020	Volatility	GARCH	COVID-19 increased volatility.	2019-2020	Global
Rakshit & Neog	2022	Exchange rate & oil price	EGARCH	The exchange rate lowered volatility, while oil prices increased it.	2009-2020	BRIC
Li et al.	2022	Stock Volatility & GDP & VIX	GARCH	COVID-19 caused a market crash.	2019-2020	Global
Fernandez-Perez et al.	2021	Volatility & stringency Index	Event Study	Cultural differences affected market reactions to COVID-19.	2020	Global
Endri et al.	2021	Stock Price Volatility & Abnormal Return	event study, GARCH	COVID-19 drove high volatility and falling stock prices in Indonesia.	2020	Global, especially Indonesia
Wang et al.	2024	Volatility	GARCH	COVID-19 triggered short- and long-term stock market volatility in Pakistan.	2020-2021	Pakistan
Papadamou et al.	2020	Google Trends & Volatility	GARCH	Google Trends regarding COVID-19 heightened stock market volatility, especially in Europe.	2019	Global
Khan et al.	2023	Financial Markets & Volatility	GARCH & EGARCH	COVID-19 volatility mostly affected Bitcoin and crude oil markets.	2018-2021	Global
Li et al.	2020	VIX, RV	HAR-USRV-VIX	In France, the UK, and Germany, there was an increase in volatility during the COVID-19 period	2000-2020	France, UK, Germany
Khan et al.	2024	Volatility	GARCH	COVID-19 sharply increased stock market volatility.	2016-2019	China, UK, Pakistan, India, US
Insaidoo et al.	2021	Return, Volatility	EGARCH	The study finds that COVID-19 slightly reduced Ghana stock returns but increased volatility by 8.23%.	2015-2020	Ghana
Yiming et al.	2024	Crude Oil, Stock Market Return, Gold, Bond	DCC-GARCH	The study finds that environmental securities hedge well against commodities and US stocks.	2017-2022	E-7 Countries
Mao et al.	2024	Crude Oil Volatility,	Diebold and Yilmaz	The study finds that COVID-19 caused prolonged	2000-2022	China, India, the United

		Stock Market Return, and Volatility	Methodology	volatility in oil and stock markets, with IDEMVT strongly influencing oil returns.		States

Empirical Analysis

Data

In this study, we used six individual sectors and one aggregate index (NASDAQ). Each company represents different sectors of the stock market: Communication Services, Technology & Information, Health Care, Financials, Consumer Discretionary, and Industrials. These specific sectors represent 85.5% of the total market cap in the United States. If the market value of one sector is less than 15% of the total, we selected 1 company, whereas if the sector made up more than 15% of the market cap, we then selected 2 companies to represent that sector.

We took into account the Starting period and ending period of COVID-19, March 11, 2020, to May 5, 2023, because on **March 11, 2020**, the WHO (World Health Organization) officially declared COVID-19 a **pandemic** until **May 5, 2023**, when the WHO declared that COVID-19 was **no longer a global health emergency**, though the virus continues to circulate.

Descriptive Statistics

Table 1. Descriptive Statistics during the COVID-19 Period(From March 11, 2020 to May 5, 2023)

Variable	Os	Mean	Std. Dev.	Min	Max
Return Nasdaq	796	0	.018	-.131	.089
Return Amazon	796	0	.025	-.151	.127
Return Apple	796	.001	.022	-.138	.113
Return Ge	796	.001	.027	-.164	.137
Return Google	796	.001	.022	-.118	.09
Return Microsoft	796	.001	.021	-.159	.133
Return United	796	.001	.02	-.19	.12
Volatility Nasdaq	796	.017	.008	.006	.052
Volatility Amazon	796	.023	.008	.009	.044
Volatility Apple	796	.021	.009	.01	.06
Volatility Ge	796	.025	.012	.012	.075
Volatility Google	796	.021	.008	.007	.05
Volatility Microsoft	796	.02	.009	.007	.062
Volatility United	796	.017	.011	.007	.071

Model

To understand the effect of the COVID-19 pandemic on 6 different sectors, we estimated the following model. In this model, the dependent variable is the volatility of individual stocks, and the independent variables are the volatility of NASDAQ and the COVID-19 dummy variable.

COVID-19 is the dummy variable that takes 1 during the pandemic from March 11, 2020, to May 5, 2023, and 0 otherwise.

$$\text{Volatility of Individual Stocks}_t = a_0 + b_1 \text{Volatility of Nasdaq}_t + b_2 \text{COVID-19}_t + \varepsilon_t$$

Findings

Figure 1. Price, Return, and Volatility of Nasdaq

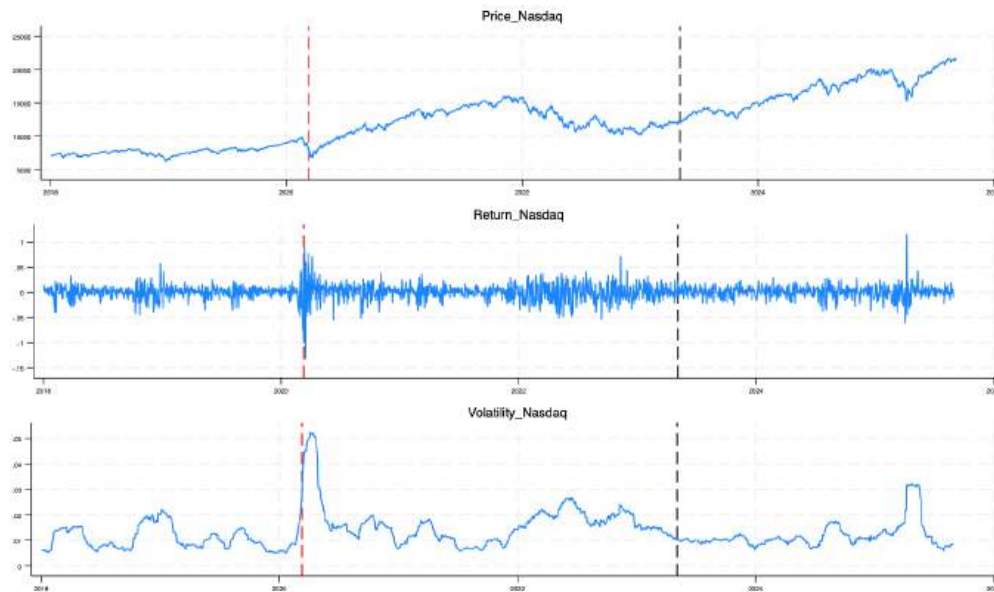


Figure 1 demonstrates the price, daily returns, and volatility of the Nasdaq index during the COVID-19 period. The red dashed line represents the start of the pandemic period, 11 March 2020, according to the World Health Organization (WHO), which declared COVID-19 a worldwide pandemic. The black dashed line represents the end of the pandemic period, 5 May 2023, declared by the WHO. The price of Nasdaq reveals a significant drop at the start of the pandemic period, reflecting the initial market shock and uncertainty among investors. The return of the Nasdaq also supports multiple spikes at the beginning of the pandemic, underlining extreme daily return changes. The spike of volatility is consistent with the start of the pandemic, indicating high market uncertainty and panic from investors. As the pandemic normalized, the volatility declined, indicating more stable market conditions as investors grew more accustomed to the new normal.

Figure 2. Price, Return, and Volatility of Amazon

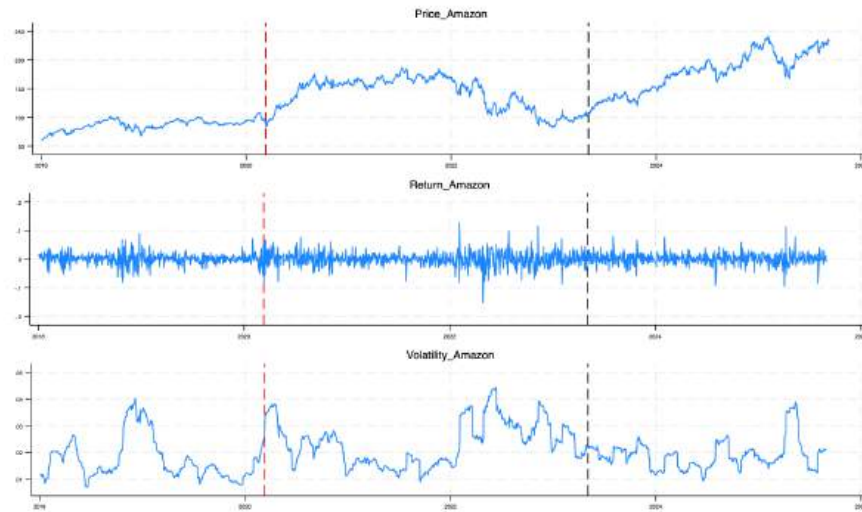


Figure 2 underlines the price, daily returns, and volatility of Amazon stock during the COVID-19 period. Amazon's price shows an initial dip in the starting period of the pandemic, reflecting investor uncertainty. Daily returns reveal sharp fluctuations in the early pandemic, with multiple spikes highlighting market instability. Volatility spiked at the start, consistent with heightened investor anxiety, but gradually declined over time, suggesting that Amazon benefited from online shopping demand during the lockdown.

Figure 3. Price, Return, and Volatility of Apple

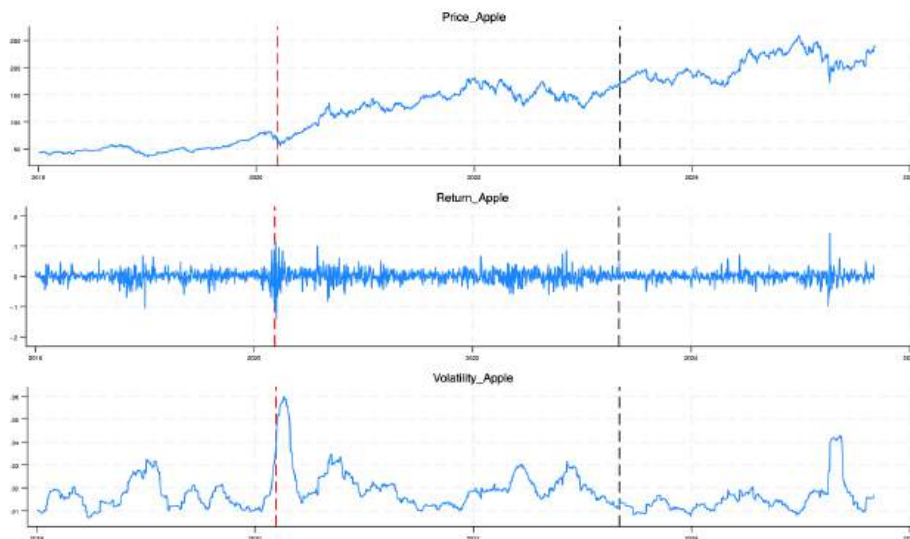


Figure 3 highlights the price, daily returns, and volatility of Apple stock during the COVID-19 period. Apple's price experienced an early dip at the onset of the pandemic, reflecting broad investor uncertainty and supply chain disruptions. Daily returns show notable spikes in the first months, capturing the market's unstable reaction to global shutdowns. Volatility rose sharply in March 2020, consistent with investor anxiety, but gradually stabilized.

Over time, Apple’s strong product ecosystem, resilient consumer demand, and expansion in digital services supported recovery, allowing its stock to grow throughout the pandemic period.

Figure 4. Price, Return, and Volatility of General Electric

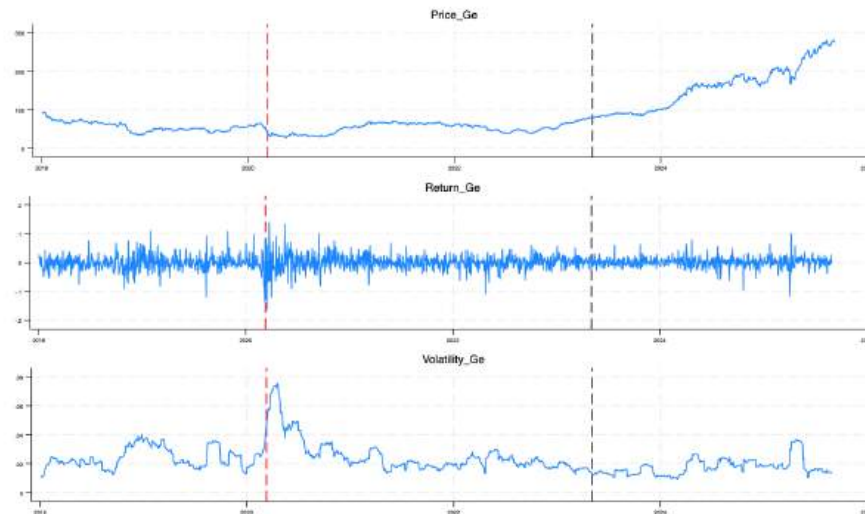


Figure 4 highlights the price, daily returns, and volatility of GE stock during the COVID-19 period. GE’s price experienced a sharp decline at the start of the pandemic. Daily returns show significant spikes in the early months, underscoring market instability. Volatility rose sharply in March 2020, consistent with investor panic, but gradually eased as market conditions stabilized. Over time, GE’s restructuring efforts and slow recovery in industrial demand helped support stabilization.

Figure 5. Price, Return, and Volatility of Google

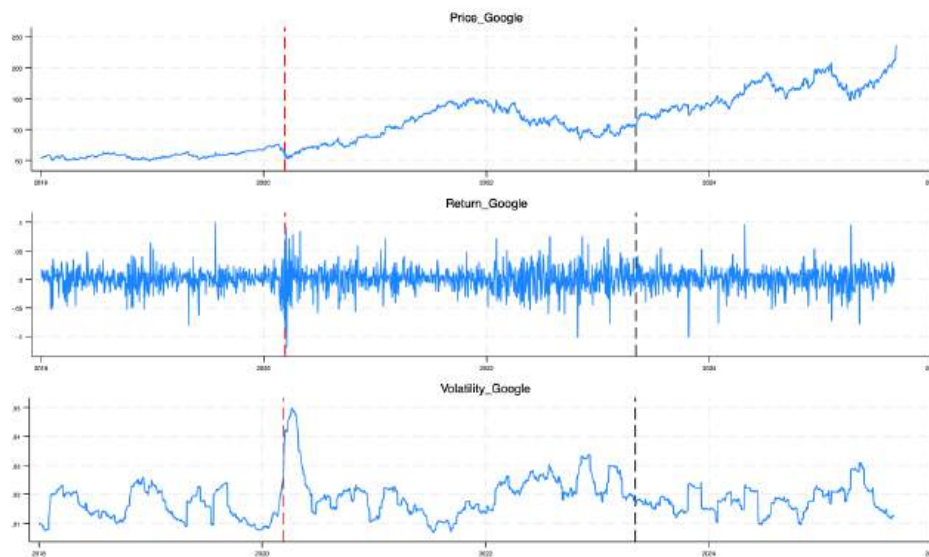


Figure 5 highlights the price, daily returns, and volatility of Google stock during the COVID-19 period. Google’s price saw an initial dip at the start of the pandemic. Daily returns

show sharp spikes in the early months, underscoring heightened market volatility. Volatility rose significantly in March 2020 but gradually declined as conditions stabilized. Over time, increased demand for online services supported Google’s recovery and allowed its stock to reach new peaks during the pandemic period.

Figure 6. Price, Return, and Volatility of Microsoft

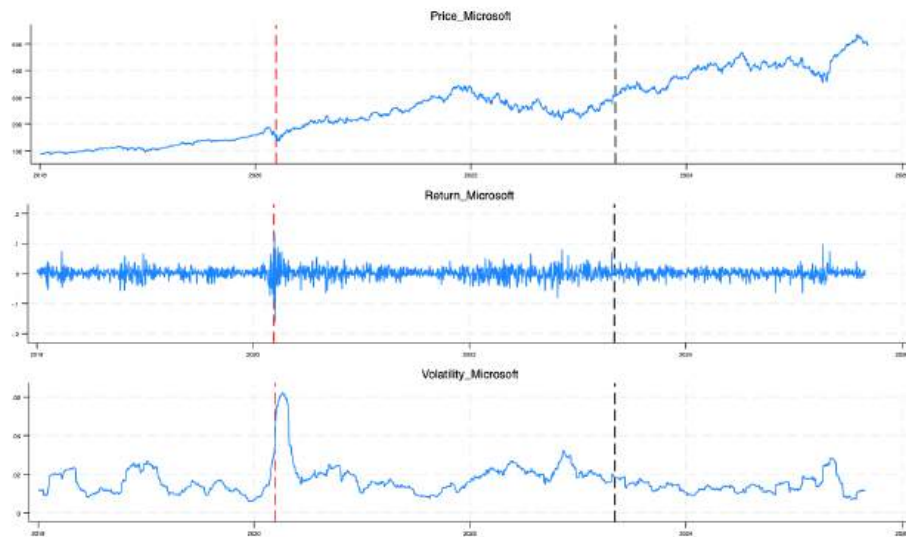


Figure 6 highlights the price, daily returns, and volatility of Microsoft stock during the COVID-19 period. Microsoft’s price experienced an initial dip at the start of the pandemic. Daily returns show sharp spikes in the early months, underscoring heightened market volatility. Volatility rose significantly in March 2020 but declined as conditions stabilized. Over time, strong demand for cloud computing, remote work solutions, and software services supported Microsoft’s recovery.

Figure 7. Price, Return, and Volatility of United Healthcare

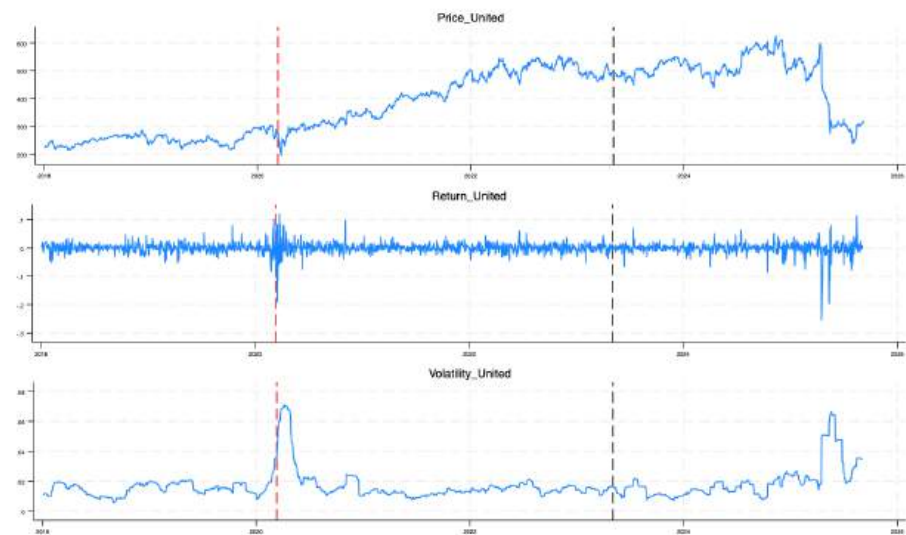


Figure 7 illustrates the price, daily returns, and volatility of UnitedHealth stock during the COVID-19 period. UnitedHealth’s price dropped at the beginning of the pandemic, reflecting

uncertainty in the healthcare sector. Daily returns show noticeable swings in the early months, highlighting unstable market conditions, but over the pandemic, the price dramatically increased. Volatility surged in March 2020 but steadily decreased as investors gained confidence. Over time, sustained demand for healthcare services and the company's diversified insurance business supported recovery and drove stabilization in UnitedHealth's stock performance.

Table 1. Regression Analysis

	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)	Model (6)
VARIABLES	V_AMZN	V_AAPL	V_GE	V_GOOG	V_MSFT	V_UNH
V_NASDAQ	0.866*** (0.036)	1.037*** (0.011)	0.997*** (0.031)	0.825*** (0.010)	1.006*** (0.019)	1.103*** (0.043)
COVID-19 Dummy Variable	0.001*** (0.000)	-0.000 (0.000)	-0.001** (0.000)	-0.000** (0.000)	0.000* (0.000)	-0.006*** (0.000)
Constant	0.008*** (0.000)	0.004*** (0.000)	0.009*** (0.000)	0.007*** (0.000)	0.003*** (0.000)	0.005*** (0.000)
Observations	1,926	1,926	1,926	1,926	1,926	1,926
R-squared	0.623	0.806	0.505	0.762	0.869	0.500

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1 Regression Analysis

Table 1 presents the results in 6 different regression models for AMZN, AAPL, GE, GOOG, MSFT, and UNH. Each column corresponds to a distinct model. If the results are statistically significant, at 1% level, we indicate by (***), the 5% level(**), and the 10% level(*). Across all models, the coefficient of the stock market index (V_NASDAQ) is positively and highly significant at 1% level, indicating that Nasdaq movements are not strongly associated with the volatility of individual stocks. The COVID-19 dummy variable captures the effect of the pandemic period on stock volatility. The COVID-19 dummy has a statistically significant and positive effect on the volatility of AMZN and MSFT, while statistically significant and negative effects on UNH, GOOG, and GE. The pattern suggests that the impact of the COVID-19 pandemic is sector-specific; it was beneficial for some and harmful to others. The companies affected negatively, like UNH, GOOG, and GE, experienced a decline in volatility during the pandemic. The investors in this company viewed these as relatively safer investments than companies like AMZN and MSFT. The COVID-19 pandemic had relatively different effects on different sectors. This emphasizes the heterogeneous impact of the pandemic among differing sectors.

Conclusion

This paper analyzed the behavior of the stock market index during the COVID-19 pandemic period: from March 11, 2020, to May 5, 2023. Using daily data from 6 key industries, the study employed regression analysis to understand how the pandemic influenced sectoral volatility in the United States' stock market. The findings demonstrate that the pandemic had a strong impact across different sectors. The results specifically indicate that the pandemic had statistically significant positive effects on the volatility of companies in the technology & information and consumer discretionary sectors. Whereas the worldwide disease had a significantly negative effect on the volatility of companies in the following sectors: healthcare, communication services, and industrials, after controlling for the effect of the aggregate stock market index (NASDAQ). This paper underlined the importance of portfolio diversification during periods of crisis.

Future research could analyze the volatility of companies outside of the main 6 that we studied in this paper. Future research could also compare findings across multiple countries' stock markets and analyze the post-pandemic period to understand long-term financial resistance.

Appendix

Appendix 1. Data Definition

Ticker Symbol	Variable Name	Sector Name	Source
NASDAQ: .IXIC	Nasdaq Inc	Aggregate Stock Market Index	Google Finance
GOOGLE: GOOG	Alphabet Inc Class C	Communication Services	Google Finance
MICROSOFT: MSFT	Microsoft Corp	Technology & Information	Google Finance
APPLE: AAPL	Apple Inc	Technology & Information	Google Finance
UNITED HEALTHCARE: UNH	United Health Group Inc	Health Care	Google Finance
AMAZON: AMZN	Amazon.com INC	Consumer Discretionary	Google Finance
GENERAL ELECTRIC: GE	General Electric Co	Industrials	Google Finance

Works Cited

- Bora, D., & Basistha, D. (2021). The outbreak of COVID-19 pandemic and its impact on stock market volatility: Evidence from a worst-affected economy. *Journal of public affairs*, 21(4), e2623.
- Endri, E., Aipama, W., Razak, A., Sari, L., & Septiano, R. (2021). Stock price volatility during the COVID-19 pandemic: The GARCH model. *Investment Management & Financial Innovations*, 18(4), 12.
- Fernandez-Perez, A., Gilbert, A., Indriawan, I., & Nguyen, N. H. (2021). COVID-19 pandemic and stock market response: A culture effect. *Journal of behavioral and experimental finance*, 29, 100454.
- Insaiddoo, M., Arthur, L., Amoako, S., & Andoh, F. K. (2021). Stock market performance and COVID-19 pandemic: evidence from a developing economy. *Journal of Chinese Economic and Foreign Trade Studies*, 14(1), 60-73.
- Khan, M. N., Fifield, S. G., & Power, D. M. (2024). The impact of the COVID 19 pandemic on stock market volatility: evidence from a selection of developed and emerging stock markets. *SN Business & Economics*, 4(6), 63.
- Khan, M., Kayani, U. N., Khan, M., Mughal, K. S., & Haseeb, M. (2023). COVID-19 pandemic & financial market volatility; evidence from GARCH models. *Journal of Risk and Financial Management*, 16(1), 50.
- Li, W., Chien, F., Waqas Kamran, H., Aldeehani, T. M., Sadiq, M., Nguyen, V. C., & Taghizadeh-Hesary, F. (2022). The nexus between COVID-19 fear and stock market volatility. *Economic research-Ekonomiska istraživanja*, 35(1), 1765-1785.
- Li, Y., Liang, C., Ma, F., & Wang, J. (2020). The role of the IDEMV in predicting European stock market volatility during the COVID-19 pandemic. *Finance research letters*, 36, 101749.
- Mao, Z., Wang, H., & Bibi, S. (2024). Crude oil volatility spillover and stock market returns across the COVID-19 pandemic and post-pandemic periods: An empirical study of China, US, and India. *Resources Policy*, 88, 104333.
- Papadamou, S., Fassas, A., Kenourgios, D., & Dimitriou, D. (2020). Direct and indirect effects of COVID-19 pandemic on implied stock market volatility: Evidence from panel data analysis.
- Rakshit, B., & Neog, Y. (2022). Effects of the COVID-19 pandemic on stock market returns and volatilities: evidence from selected emerging economies. *Studies in Economics and Finance*, 39(4), 549-571.
- Uddin, M., Chowdhury, A., Anderson, K., & Chaudhuri, K. (2021). The effect of COVID-19 pandemic on global stock market volatility: can economic strength help to manage the uncertainty?. *Journal of Business Research*, 128, 31.
- Wang, B., Waris, M., Adamiak, K., Adnan, M., Hamad, H. A., & Bhatti, S. M. (2024). The effects of the COVID-19 pandemic period on stock market return and volatility. Evidence from the Pakistan Stock Exchange. *Plos one*, 19(4), e0295853.
- Yiming, W., Xun, L., Umair, M., & Aizhan, A. (2024). COVID-19 and the transformation of emerging economies: financialization, green bonds, and stock market volatility. *Resources Policy*, 92, 104963.