IJHSR International Journal of High School Research

August 2025 | Volume 7 | Issue 8
ijhsr.terrajournals.org
ISSN (Print) 2642-1046
ISSN (Online) 2642-1054

Marine Biology Research at Bahamas

Unique and exclusive partnership with the Gerace Research Center (GRC) in San Salvador, Bahamas to offer marine biology research opportunities for high school teachers and students.

- Terra has exclusive rights to offer the program to high school teachers and students around world.
- All trips entail extensive snorkeling in Bahamian reefs as well as other scientific and cultural activities.
- Terra will schedule the program with GRC and book the flights from US to the GRC site.
- Fees include travel within the US to Island, lodging, meals, and hotels for transfers, and courses.
- For more information, please visit terraed.org/bahamas.html

Terra is a N.Y. based 501.c.3 non-profit organization
dedicated for improving K-16 education

Table of Contents

August 2025| Volume 7| Issue 8

Editorial Board

International Journal of High School Research

IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Risk-Return Dynamics of Renewable vs Non-Renewable Energy Portfolios: An Analysis of Volatility and Returns

Jeremy Cheung

Reading School, Erleigh Road, Reading, Berkshire, RG1 5LW, UK; jeremycheung2007@gmail.com

ABSTRACT: This paper examines the performance dynamics of renewable and non-renewable energy portfolios by analyzing the relationship between volatility and returns. Using a dataset of daily closing prices from 2021 to 2024, covering eight stocks each in the renewable energy and non-renewable energy sectors, and the NASDAQ Composite Market Index, the study identifies a distinct risk-return trade-off among all benchmark portfolios. The findings show that the non-renewable energy portfolio exhibits higher returns but also increased volatility, highlighting that greater risk is associated with greater returns. In contrast, the renewable energy portfolio demonstrates the highest average volatility while producing relatively low returns, suggesting it may not be an efficient investment choice. However, the diversified composite energy portfolio successfully generated higher returns than the market while offering lower risk in terms of volatility. The paper concludes that investors in the energy sector must navigate different risk preferences: those seeking higher returns may be drawn to non-renewable energy, while those prioritizing stability in an evolving energy market may prefer to diversify their portfolios with some renewable energy stocks. These insights offer valuable information for portfolio construction and risk management strategies.

KEYWORDS: Mathematics, Analysis, Volatility, Cumulative Return, Renewable and Non-renewable Energy.

Introduction

In recent years, the renewable energy sector has grown in popularity and size within the world of investment, driven by increasing awareness of climate change and the transition to sustainable energy production. However, investors must navigate the unique complexities of risk-return relationships when investing in this field. This study aims to analyse the volatility and return of five distinct portfolios, composed of renewable energy, non-renewable energy, the market index, a composite energy portfolio, and a combined portfolio.

A consensus in the existing literature states that risk and return form a direct, positive correlation. An investment with higher risks generally requires higher returns for it to be considered reasonable and efficient.[1] This fundamental relationship underpins modern portfolio theory, as introduced by Markowitz.[2]

Risk in the context of investment is defined as the potential for the actual returns to differ from expected returns.[3] This divergence typically involves the possibility of losing some or all of the original investment, while also referring to the variability of returns around the expected outcome. Although risk is difficult to quantify, it can be assessed using several indicators.

Firstly, the uncertainty of returns is captured by volatility, which measures stock price fluctuations over time. Higher volatility typically indicates greater risk. Standard deviation provides a similar measure of dispersion, indicating the extent to which returns deviate from their mean. This addresses the fundamental characteristics of risk.

Risk can be further categorized into systematic and unsystematic risk. Systematic risk, also known as market risk, arises from broader economic variables such as recessions, inflation, and changes in interest rates. This cannot be mitigated effectively through diversification as its impact radiates throughout the entire market, though in varying degrees. Unsystematic risk, on the other hand, is company or industry-specific and can be reduced through diversification.[4]

This study provides a comprehensive analysis of the volatility and returns of renewable and non-renewable energy companies against a market index. We have also built a composite energy portfolio and a customized combined portfolio to assess the role of diversification. Using statistical techniques such as t-tests, scatter plots, and regression analysis, this research contributes to the growing literature on energy sector integration and portfolio optimization. These insights are particularly relevant for investors seeking energy-focused diversification strategies.[5]

Under the current political climate, renewable energy sources are expected to gradually replace fossil fuels. However, recent geopolitical conflicts have injected uncertainty into global energy markets. The economic sanctions and destruction of infrastructure during the war in Ukraine have disrupted global energy supply chains, especially between Russia and Europe. Consequently, the growing demand for energy independence reinforced short-term reliance on traditional fossil fuels while uncovering renewable energy's costly and unreliable nature, as well as their lack of technological maturity.[6] Integrated oil and gas companies had benefited from increased energy prices, leading to significant revenue growth. Investor sentiments, therefore, shifted towards traditional energy stocks, whose stability is represented by strong cash flows, regular dividend payments, and share buybacks.[7]

This paper reassesses the relationship between renewable and non-renewable energy investments using a recent dataset. It examines whether one investment strategy outperforms the

DOI: 10.36838/v7i8.1

other given the current geopolitical landscape. The study's key contributions include an empirical analysis of volatility and returns in renewable and non-renewable energy sectors, offering investors guidance on balancing risk and optimizing returns.

This paper proceeds as follows: Section 2 provides background information on the existing literature, Section 3 highlights the empirical findings, and Section 4 concludes the paper.

Literature Review

The global energy market has undergone substantial changes during the past several years, with the increasing participation of renewable energy in addition to conventional fossil fuels. This shift has introduced new dynamics in energy markets and investment portfolios that require greater insight into market integration, portfolio optimization, and risk management strategies.[8]

Recent studies have proved that renewable energy and conventional energy markets are highly connected. Zhang *et al.*[9] emphasize that renewable energy equities demonstrate a significant correlation with the returns of fossil energy under extreme market conditions, yet this relationship diminishes under normal market conditions. Xia *et al.*[10] also confirm the results by exhibiting asymmetric and significant impacts of energy price volatility on the return of renewable energy firms, particularly in European markets.

Such a relationship differs by various market conditions. Li *et al.*[11] discovered a positive correlation between the renewable energy and fossil energy markets in normal market conditions. In bear markets, however, it becomes extreme and contains strong asymmetrical dynamics. Jiang *et al.*[12] built on this by demonstrating that renewable energy shares have a net positive effect on the fossil energy markets, particularly in the oil and coal markets, while the effect is also highly periodic in the gas market.

Market integration patterns show wider variations in geographical terms. Bianconi and Yoshino[13] analysed 64 oil and gas companies in 24 nations, discovering that specific and common risk factors also pose a substantial influence on stock returns. Firm size and leverage were highlighted as key factors by the research, especially after the 2008 financial crisis.

Valadkhani[14] discovered that renewable energy ETFs outperformed fossil fuel ETFs in the US market, especially in risk-adjusted performance measures such as the Sortino and Sharpe ratios. According to the VIX index, the performance difference is more significant when the market uncertainty is greater.

The development of portfolio optimization techniques in energy markets is seen to reflect growing sophistication in methodology and approach. Kuang illustrated that although clean energy equities underperform overall equities, they outperform fossil fuel assets on a risk-adjusted return basis. The research also established that adding clean energy to a traditional asset enhances portfolio performance.

For the Chinese market, Bai *et al.*[15] proposed an enhanced portfolio approach that surpassed classical Markowitz approaches under varying market conditions. This was subsequently confirmed by Ma *et al.*[16]

Research has increasingly focused on energy portfolio risk management. Ahmad[17] found that crude oil serves as a better hedge for clean energy stocks than for technology stocks, particularly during periods of crisis. Galvani and Plourde[18] demonstrated that while energy futures reduce portfolio risk, they offer no significant improvement in energy stock returns.

Wang *et al.*[19] carried this research forward to commodity futures, discovering substantial gains from employing energy futures in portfolio diversification, especially within commodity portfolios.

The impact of policy decisions on renewable energy markets has been extensively documented. Antoniuk and Leirvik[20] discovered that policy events related to climate change have significantly influenced market returns. According to the research, policymakers must consider the reaction of the stock market to climate risk since investors quickly respond to climate news.

According to Masini and Menichetti,[21] investors trust mature technology more than policy intervention and are greatly affected by external advisers and peer pressure.

The advent of renewable energy markets can be encapsulated by changing investment patterns. Nautiyal *et al.*[22] discovered that energy-weighted portfolios have the most potential to provide the best returns in the short term, particularly in volatile times such as the global financial crisis and COVID-19. The green equities were determined by their study to be effective in hedging and risk management.

In emerging economies, the studies report controversial evidence. Artini and Sandhi[23] compared SME and manufacturing stock portfolios in Indonesia, India, and China and reported higher performance in Chinese and Indian markets compared to the Indonesian market. The geographical disparity in performance suggests the role of market-specific determinants in portfolio selection.

The literature has also considered the contribution of institutional drivers to the performance of energy markets. Antônio *et al.*[24] recommended careful consideration of market data. The results are consistent with Shachmurove's[25] previous research on Latin American markets, which called for careful consideration of risks and opportunities specific to each market.

There is recent proof of diverse methodological trends. Dai *et al.*[26] applied TVP-VAR methods in the investigation of volatility spillover among crude oil, gold, and Chinese new energy markets. Wang *et al.*[27] applied network analysis in dynamic spillover comprehension in the energy stock market, whereas Gurrib *et al.*[28] applied cryptocurrency analysis in energy portfolio optimization. The contrast of research methods indicates an ongoing enhancement in theoretical comprehension as well as empirical applications.

Shrimali[29] and Gargallo *et al.*[30] are of the opinion that the effectiveness of market integration and portfolio optimization will rely on building more robust policy platforms and risk management instruments. This will be critical in achieving investment targets and wider sustainability objectives.

Methods

The paper has used descriptive statistics, scatter plot graphics, t-tests, and regression analysis.

Pairwise T-test:

The t-test aims to evaluate the null hypothesis (H_0), which typically shows that there is no difference between the means of the two groups being compared. The alternative hypothesis (H_1) implies that there is a significant difference.

H0: $\mu 1 = \mu 2$

H1: $\mu 1 \neq \mu 2$

The test statistics for a t-test are calculated using formula[31]

$$t = \frac{x1 - x2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Regression Analysis:

The model offers a practical framework for evaluating investment stability by capturing how returns respond to shifts in volatility. Volatility serves as a proxy for financial risk, enabling the examination of how different stock portfolios influence market uncertainty, hence chosen as the dependent variable. This provides insights into sector-specific drivers of volatility and potential volatility spillovers across markets.

$Yt = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{1t} + \ldots + \beta_n X_{nt} + \varepsilon_t$

Where Y_t is the dependent variable for observation t, which refers to volatility.

β_0 is the constant term representing the expected value of the dependent variable when all independent variables are 0.

β_1 to β_n are coefficients for the independent variables, which include returns of renewables, non-renewables, the aggregate stock market index, composite energy portfolio, and the combined portfolio.

While keeping all other variables constant, each coefficient shows how much the dependent variable changes when the corresponding variable changes by 1 unit.

ε_t is the error term, which represents the difference between the actual value and the predicted value from the model.

Volatility Calculations:

The volatility of each financial instrument is measured by employing the 30-day rolling standard deviation of daily returns.

Portfolio Construction:

The 5 portfolios are constructed as follows:

Portfolio X represents equally weighted averages of the eight individual renewable stocks for both cumulative return and volatility.

Portfolio X = 1/8*(NEE + CWEN + HASI +NEXNY + BEP + FLNC + ADANIGREENNS + FSLR)

Portfolio Y represents equally weighted averages of the eight individual non-renewable stocks for both cumulative return and volatility.

Portfolio Y = 1/8*(XOM + CVX + PCCYF + SHEL + TTE + COP + BP + EQNR)

Portfolio Z represents the aggregate market index taken directly from the NASDAQ Composite index for both cumulative return and volatility.

Portfolio XY is the equally weighted combination of Portfolio X and Portfolio Y.

Portfolio XY = 1/2*(Portfolio X + Portfolio Y)

Portfolio XYZ is the equally weighted combination of Portfolios X, Y, and Z.

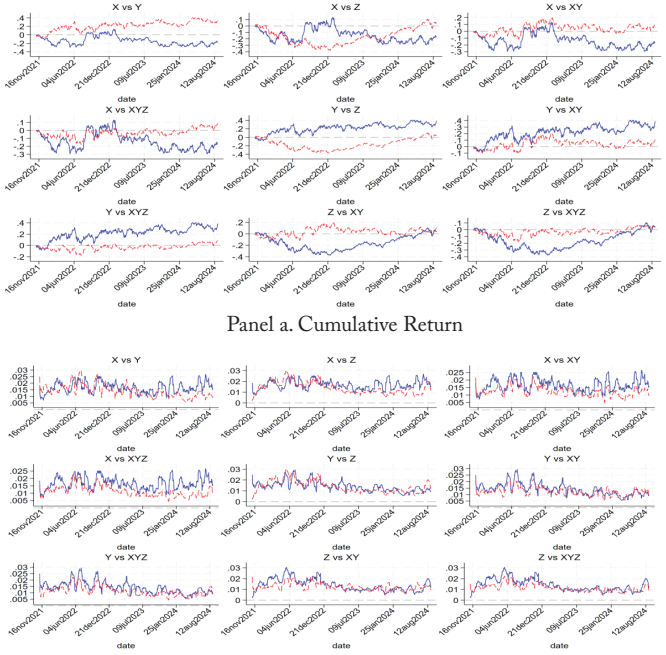
Portfolio XYZ = 1/3*(Portfolio X + Portfolio Y+Portfolio Z)

Result and Discussion

Data:

The data has been obtained from Yahoo Finance for a selected bundle of renewable energy companies (ticker symbols: NEE, CWEN, HASI, NEXNY, BEP, FLNC, ADANIGREENNS, and FSLR), non-renewable companies (ticker symbols: XOM, CVX, PCCYF, SHEL, TTE, BP, COP, and EQNR), along with the stock market index- NASDAQ Composite (ticker symbol: IXIC). The data covers the daily closing price for the selected period from 1/11/2021 to 30/8/2024. Within the selected period, there were 712 observations for all individual assets.

Empirical Findings:



Panel a. Cumulative Return

Panel b. Volatility

Figure 1: The line graphs show paired comparisons of changes in cumulative return (Panel a) and volatility (Panel b) for benchmark portfolios throughout the observation period. Both variables created asymmetrical patterns among portfolios, particularly in terms of cumulative returns.

Figure 1 presents the line graphics of cumulative return and volatility in Panel a and Panel b, respectively. In Panel a, we compared the cumulative return of portfolios X, Y, Z, XY, and XYZ. The order of average cumulative returns from best to worst is Y, XY, XYZ, X, and Z. Moreover, the average cumulative return of Y is much higher than the rest. Portfolio X did outperform the market significantly during the period from

DOI: 10.36838/v7i8.1

June to December 2022 before it started to decline and finally fell below market returns. In Panel b, we compared the volatility of portfolios X, Y, Z, XY, and XYZ. The order of ascending average volatility is XYZ, XY, Z, Y, and X. An asymmetry is observed in the risk-return relationship: portfolio X ranks 4th in cumulative return while portfolio Y ranks 1st. However, portfolio X is more volatile than Y, suggesting unfavorable risk-return dynamics, thus a less efficient investment choice.

Table 1 provides descriptive statistics for various cumulative return variables for Portfolio X, Y, Z, XY, and XYZ. This table includes the number of observations, mean return, standard deviation, and the maximum and minimum value of return. Several observations can be drawn from the comparisons of all portfolios. Portfolio Z, representing the market index, produced the lowest mean cumulative return, primarily due to a period of significant negative returns as indicated by a minimum value of -0.38. Although having a slightly higher maximum value in return, its mean return remains marginally lower than that of portfolio X. Portfolio Y demonstrates a dominant performance with the highest minimum, maximum, and mean value. This aligns with the positive impacts experienced by oil and gas companies, contrasted with the negative effects felt by the broader market, particularly the non-renewable energy sector, during the energy crisis and the Ukraine war. Moreover, we can see that diversification hugely decreases the magnitude of negative returns by more than 0.1 for both portfolios compared to the market itself. Positive returns are also mostly preserved as maximums remain closer to the market value. The significance of diversification is evident in the production of a notably higher mean return compared to the original market average.

Table 1: Includes the descriptive statistics of cumulative return for all portfolios.

Variable	Obs	Mean	Std. Dev.	Min	Max
cumulative return x	712	-.142	.102	-.297	.103
cumulative return y	712	.218	.107	-.085	.408
cumulative return z	712	-.164	.123	-.38	.108
cumulative return xy	712	.038	.063	-.118	.198
cumulative return xyz	712	-.030	.049	-.179	.088

Table 2 provides the pairwise t-test analysis between the five portfolios. According to the pairwise t-test analysis, all comparisons are statistically significant at the 1 percentage level. This suggests that the patterns in cumulative return among all individual comparisons demonstrate statistical differences. Therefore, cumulative return is an important indicator that distinguishes renewable energy stocks from non-renewable energy stocks.

Table 2: Presents the Paired t-test results of cumulative return between the five benchmark portfolios

Comparisons	Observations	Mean	Standard Error	Standard deviation	Base (mean)	Diff	t value	p value
vs C_X								
C_Y	712	.217799	.0039931	.1065486	-.1423138	-.347	-58.1	**0.000**
C_Z	712	-.1644995	.004624	.1233836	-.1423138	.036	3.1	**0.002**
C_{XY}	712	.0377426	.0023743	.0633533	-.1423138	-.174	-58.1	**0.000**
C_{XYZ}	712	-.0296714	.0018455	.0492446	-.1423138	-.105	-28.1	**0.000**
vs C_Y								
C_Z	712	-.1644995	.004624	.1233836	.217799	.039	64.6	**0.000**
C_{XY}	712	.0377426	.0023743	.0633533	.217799	.186	58.1	**0.000**
C_{XYZ}	712	-.0296714	.0018455	.0492446	.217799	.254	76.03	**0.000**
vs C_Z								
C_{XY}	712	.0377426	.0023743	.0633533	-.1644995	-.191	-34.9	**0.000**
C_{XYZ}	712	-.0296714	.0018455	.0492446	-.1644995	-.127	-34.9	**0.000**

Table 3 provides descriptive statistics for various volatility variables used in the analysis. The volatility statistics contrast sharply with the trends observed in cumulative returns. Portfolio X offers the highest mean volatility, indicating the largest fluctuations in stock prices. With reference to our previous cumulative return figures, we can infer that it experiences more frequent and pronounced downward price movements compared to other portfolios. This signals a loss of confidence among investors, especially in the renewable energy sector, in reaction to the crises. Despite a huge divergence in returns, portfolios Y and Z offer similar volatilities. This finding suggests that portfolio Y may be a more attractive investment, assuming that volatility is accepted as an accurate measure of risk. As expected, the diversified portfolios (XY & XYZ) have produced significantly lower mean volatility values compared to other benchmark portfolios. Additionally, the reduction in overall portfolio volatility in further diversification shows that renewable energy, non-renewable energy, and the market index are not perfectly correlated assets.

Table 3: Includes the descriptive statistics of volatility for all portfolios.

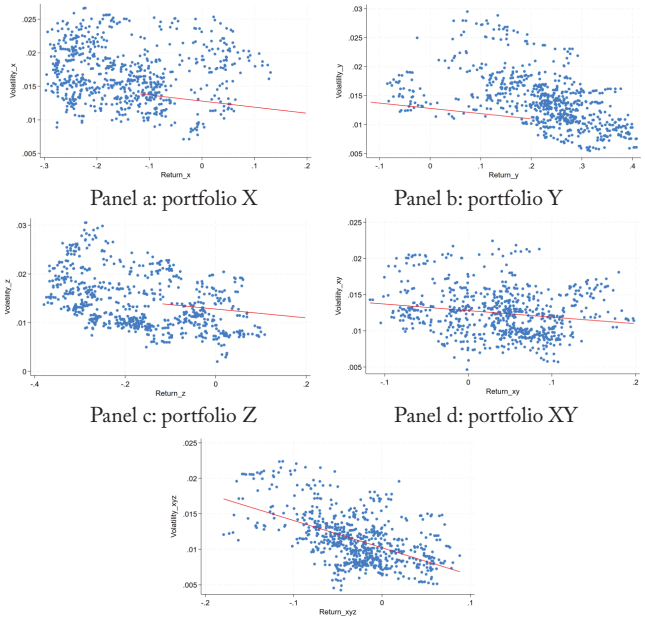
Variable	Obs	Mean	Std. Dev.	Min	Max
volatility x	711	.0167	.0041	.0071	.0267
volatility y	711	.0141	.0045	.0055	.0295
volatility z	711	.014	.0055	.002	.0306
volatility xy	711	.0124	.0032	.0046	.0224
volatility xyz	711	.0114	.0036	.0042	.0224

Table 4 provides the pairwise t-test analysis between the five portfolios. According to the pairwise t-test analysis, all comparisons are statistically significant at the 1 percentage level except the comparison between portfolio Y & portfolio Z. This suggests that the patterns in volatility among most individual comparisons are also statistically different. Therefore, like cumulative return, volatility is another important indicator that distinguishes renewable energy stocks from non-renewable energy stocks.

Table 4: Present the Paired t-test results of volatility between the five benchmark portfolios.

Comparison	Observations	Mean	Standard Error	Standard deviation	Base (Mean)	Diff	t value	p value
vs V_X								
V_Y	711	.0140936	.0001704	.0045449	.0167218	.0030201	13.1644	**0.0000**
V_Z	711	.0140433	.000206	.0054931	.0167218	.0030726	13.3459	**0.0000**
V_{XY}	711	.01245	.0001201	.0032017	.0167218	.0044918	38.1254	**0.0000**
V_{XYZ}	711	.0113569	.0001341	.0035759	.0167218	.0056245	40.5681	**0.0000**
vs V_Y								
V_Z	711	.0140433	.000206	.0054931	.0004634	.0003872	0.2937	0.3845
V_{XY}	711	.01245	.0001201	.0032017	.0004634	.0018541	15.3357	**0.0000**
V_{XYZ}	711	.0113569	.0001341	.0035759	.0004634	.0029519	24.9652	**0.0000**
vs V_Z								
V_{XY}	711	.01245	.0001201	.0032017	.0140433	.0019109	9.8479	**0.0000**
V_{XYZ}	711	.0113569	.0001341	.0035759	.0140433	.0029127	23.2979	**0.0000**

The Relationship between risk and return:



Panel a: portfolio X

Panel b: portfolio Y

Panel c: portfolio Z

Panel d: portfolio XY

Panel e: portfolio XYZ

Figure 1: The scatter plot graphic between volatility and return shows an overall negative correlation between the two variables, implying unconventional risk-return dynamics.

Figure 2 illustrates the relationship between volatility and cumulative return for portfolios X, Y, Z, XY, and XYZ. Results indicate a negative relationship between volatility and cumulative return for all cases. A strongly negative correlation is seen in portfolio XYZ, whereas portfolios X, Y, Z, and XY show a slight negative correlation. They may not imply that higher volatility is associated with higher returns. The observed negative relationship can be attributed to the heightened economic and geopolitical uncertainty during the energy crisis and the Ukraine War. Risk-averse sentiments are more common in periods of crisis, prompting widespread selloffs across sectors. This behavior may increase market volatility while simultaneously driving down returns, thereby resulting in the temporary appearance of a negative risk-return dynamic. It may guide investors in their portfolio construction and risk management strategies. Understanding that increased volatility does not always lead to higher returns may discourage energy investors from taking more risk in hopes of greater returns.

The regression analysis in Table 5 indicates that the cumulative return has a consistently negative and statistically significant impact on the volatility of portfolios X, Y, Z, XY, and XYZ, with p-values below 0.05 across all models. The maximum impact of return on volatility can be seen for the combined portfolio with the largest coefficient of -0.0384.

Table 5: Presents the regression analysis between volatility and returns among all portfolios.

VARIABLES	(1) Volatility_x	(2) Volatility_y	(3) Volatility_z	(4) Volatility_xy	(5) Volatility_xyz
Return_x	-0.0039** (0.0015)				
Return_y		-0.0219*** (0.0016)			
Return_z			-0.0198*** (0.0014)		
Return_xy				-0.0090*** (0.0019)	
Return_xyz					-0.0384*** (0.0026)
Constant	0.0162*** (0.0003)	0.0189*** (0.0004)	0.0108*** (0.0003)	-0.0006*** (0.0001)	0.0102*** (0.0001)
Observations	711	711	711	711	711
R-squared	0.0091	0.2619	0.1974	0.0320	0.2798

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Evidence from individual companies:

To further understand the relationship between risk and return and to provide more robust empirical evidence, we have attained additional cumulative return and volatility data from individual companies, as presented in Tables 6 and 7.

Table 6: Includes the descriptive statistics for various cumulative return variables used in the analysis.

Variable	Obs	Mean	Std. Dev.	Min	Max
Panel A: Renewable Energy Companies					
cumulative return NEE	712	-.18	.128	-.468	.088
Cumulative return CWEN	712	-.22	.157	-.524	.102
cumulative return HASI	712	-.557	.194	-.83	.047
Cumulative return NEXNY	712	-.185	.11	-.446	.066
cumulative return BEP	712	-.306	.148	-.559	.000
cumulative return FLNC	712	-.694	.177	-.868	.019
cumulative return ADG~S	712	.34	.519	-.522	1.695
cumulative return FSLR	712	.018	.292	-.524	.791
Panel B: Non-Renewable Energy Companies					
cumulative return XOM	712	.426	.185	-.103	.689
cumulative return CVX	712	.282	.124	-.024	.556
cumulative return PCCYF	712	.036	.23	-.334	.656
cumulative return SHEL	712	.207	.13	-.103	.523
cumulative return TTE	712	.118	.109	-.155	.372
cumulative return COP	712	.302	.141	-.088	.653
cumulative return BP	712	.079	.093	-.155	.295
cumulative return EQNR	712	.08	.158	-.19	.502

Table 7: Includes the descriptive statistics for various volatility variables from the 16 individual company stocks used in the analysis.

Variable	Obs	Mean	Std. Dev.	Min	Max
Panel A: Renewable Energy Companies					
volatility NEE	711	.0166	.0065	.0029	.0404
volatility CWEN	711	.0188	.006	.0104	.0434
volatility HASI	711	.0328	.0115	.0006	.0674
volatility NEXNY	711	.0234	.0132	0	.0674
volatility BEP	711	.0186	.0058	.0072	.0362
volatility FLNC	711	.0518	.015	.0192	.1025
volatility ADG~S	711	.0298	.0156	.0065	.0824
volatility FSLR	711	.0301	.01	.011	.0703
Panel B: Non-Renewable Energy Companies					
volatility XOM	711	.0169	.0052	.007	.0331
volatility CVX	711	.0155	.0055	.0041	.0315
volatility PCCYF	711	.024	.0099	0	.0701
volatility SHEL	711	.0154	.0065	.0052	.0345
volatility TTE	711	.0161	.006	.0014	.0387
volatility COP	711	.0194	.0076	.0065	.0481
volatility BP	711	.0171	.0068	.0058	.0451
volatility EQNR	711	.0205	.0058	.0079	.0362

Risk-return dynamics in individual companies:

From the individual company figures on volatility and cumulative returns, we gain detailed insights into the risk and return patterns within renewable and non-renewable energy stocks. In terms of risk, non-renewable energy stocks (XOM, CVX, PCCYF, SHEL, TTE, COP, BP, EQNR) generally exhibit lower volatility with standard deviations of around 0.004 to 0.005, suggesting greater price stability. On the other hand, renewable stocks show higher volatility individually; three

DOI: 10.36838/v7i8.1

companies (HASI, FLNC, ADANIGREENNS) have a standard deviation exceeding 0.01, indicating higher risk and more frequent price swings.

The risk-return relationship is significantly different between renewable and non-renewable energy companies. Non-renewable stocks in general offer higher returns with lower risk, reflecting consistent performance and predictable growth. In contrast, renewable energy stocks, despite higher volatility, have mostly negative returns. An anomaly is ADANIGREENNS, a renewable energy stock that demonstrates both high volatility and high returns.

For brevity, we cannot include all graphics and pairwise t-tests for individual companies against the portfolios. However, all empirical evidence is available on request.

Conclusion

The paper investigates the performance of renewable and non-renewable energy portfolios by examining volatility and cumulative returns. The findings highlight that portfolio Y, consisting of eight non-renewable energy companies, demonstrates a higher cumulative return compared to the renewable energy portfolio X and the broader market index portfolio Z. However, the higher return comes with increased volatility, indicating greater risk associated with non-renewable energy investment. In contrast, renewable energy stocks, represented by portfolio X, show the highest volatility despite producing similar returns to the market.

The results suggest that investors face distinct risk-return trade-offs when investing in renewable vs non-renewable energy sectors. Non-renewable energy stocks may appeal to investors seeking higher returns and are willing and able to tolerate higher risks. Renewable energy appears to be less efficient under the traditional risk-return framework. However, the perception is skewed to an extent by the data period, during which the energy crisis disproportionately benefited oil and gas producers. Despite this, the analysis implies that renewable energy investment could play a valuable role in diversifying portfolios.

However, several limitations should be acknowledged. The analysis is based on a relatively short time frame (2021-2024), which may limit the generalizability of the conclusions to longer-term market dynamics. Additionally, the selection of representative stocks is limited in both range and number, potentially omitting important variations across the energy sector. More comprehensive research would consider renewable energy production methods beyond solar and wind power, such as hydroelectric, geothermal, and bioenergy.

Looking ahead, a changing geopolitical landscape will continue to play a decisive role in shaping future energy markets. The Ukraine war, growing tensions in the Middle East, and concerns over supply chain dependencies have revealed strategic vulnerabilities of fossil fuel-dominated energy systems. In response, many countries are accelerating the transition toward domestic renewable energy production to achieve energy security. Such geopolitical considerations will increasingly shift capital allocation to favor more diversified and resilient portfolios with significant portions of clean energy assets. This topic continues to be an intersection of energy policy, market volatility, and global politics, deserving of future attention in academic research and investment strategies.

Future research might explore the longer-term performance of this portfolio as the renewable energy sector matures and global policies shift towards greener initiatives, as well as incorporating ESG factors as elements of risk and performance. This study provides a basis for understanding how geographical, economic, and environmental changes impact the performance of energy sector stocks, guiding investors in their portfolio allocation decisions within the energy market.

Acknowledgments

I would like to thank my mentor, Dr. Abdullah Yalaman, for guiding me through the process of conducting this research.

References

1. Sharpe, W. F. Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. *J. Finance* **1964**, *19* (3), 425–442.
2. Markowitz, H. Portfolio Selection. *J. Finance* **1952**, *7* (1), 77–91.
3. Jensen, M. C. Risk, the Pricing of Capital Assets, and the Evaluation of Investment Portfolios. *J. Bus.* **1969**, *42* (2), 167–247.
4. Ferson, W.; Schadt, R. W. Measuring Fund Strategy and Performance in Changing Economic Conditions. *J. Finance* **1996**, *51* (2), 425–461.
5. Beja, A. On Systematic and Unsystematic Components of Financial Risk. *J. Finance* **1972**, *27*, 37–45.
6. Pindyck, R. S. Volatility in Natural Resource Prices: A Review. J. *Environ. Econ. Manage*. **1999**, *37* (1), 128–150.
7. Guay, W.; Harford, J. The Cash-Flow Permanence and Information Content of Dividend Increases versus Repurchases. *J. Financ. Econ*. **2000**, *57* (3), 385–415.
8. Kuang, P. Clean Energy Investment and Portfolio Diversification. *J. Sustainable Finance Invest.* **2021**, *11* (3), 234–252.
9. Zhang, Y.; Sun, X.; Chen, W. Renewable Energy Market Integration: Evidence from Global Equity Markets. *Energy Econ*. **2023**, *120*, 106564.
10. Xia, Y.; Wang, Y.; Zhang, R. Energy Price Volatility and Renewable Energy Investment: An Asymmetric Analysis. *Econ. Model*. **2019**, *84*, 88–102.
11. Li, J.; Tang, X.; Liu, P. Fossil Fuel and Renewable Energy Markets: A Behavioral Perspective. *Energy Res. Soc. Sci*. **2022**, *91*, 102768.
12. Jiang, H.; Wu, Y.; Zhao, L. The Interplay of Renewable and Non-Renewable Energy Stocks: A Global Perspective. *J. Energy Mark*. **2021**, *14* (2), 75–96.
13. Bianconi, M.; Yoshino, J. A. Risk and Return in the Global Energy Sector. *Int. Rev. Econ. Finance* **2014**, *34*, 99–112.
14. Valadkhani, A. The Performance of Renewable Energy ETFs in the US Market. *Renew. Sustain. Energy Rev*. **2024**, *179*, 113459.
15. Bai, X.; Zhang, Y.; Liu, W. A Robust Portfolio Approach for the Chinese Energy Market. *J. Financ. Quant. Anal*. **2019**, *54* (2), 419–442.
16. Ma, F.; Li, T.; Zhao, Y. Revisiting Portfolio Optimization in Energy Markets. *Energy Econ*. **2023**, *123*, 107654.
17. Ahmad, W. Oil as a Hedge for Renewable Energy Stocks: A Time-Varying Analysis. *Appl. Energy* **2017**, *207*, 27–37.
18. Galvani, V.; Plourde, A. Futures Markets and Energy Stock Performance. *J. Futures Mark*. **2010**, *30* (5), 450–475.
19. Wang, Z.; He, K.; Liu, J. Commodity Futures and Portfolio Diversification. *J. Commod. Mark*. **2022**, *25*, 100182.

20. Antoniuk, Y.; Leirvik, T. Stock Market Responses to Climate Policy Shocks. *J. Sustainable Finance Invest.* **2024**, *14* (1), 109–125.
21. Masini, A.; Menichetti, E. Investor Preferences in Renewable Energy Markets. *Energy Policy* **2013**, *56*, 1–12.
22. Nautiyal, S.; Kaushik, G.; Singh, R. Energy Portfolios and Market Shocks: Evidence from the Global Financial Crisis and COVID-19. *J. Energy Finance Dev*. **2024**, *18*, 102345.
23. Artini, A.; Sandhi, N. Comparative Analysis of SME and Manufacturing Stocks in Emerging Markets. *J. Emerg. Mark*. **2020**, *14* (3), 257–273.
24. Antônio, D. R.; Almeida, R. P.; Lopes, A. Market Information and Energy Sector Performance. *J. Energy Econ.* **2015**, *48*, 356–368.
25. Shachmurove, Y. Latin American Energy Markets: Risks and Opportunities. *J. Int. Financ. Mark. Inst. Money* **1998**, *8* (3–4), 343–359.
26. Dai, J.; Wang, P.; Liu, X. TVP-VAR Analysis of Energy Market Spillovers. *Energy Econ*. **2022**, *111*, 106569.
27. Wang, Y.; Guo, Z.; Zhang, Q. Network Analysis of Energy Stock Spillovers. *Energy Policy* **2020**, *139*, 111252.
28. Gurrib, I.; Kamalov, F.; Rezgui, S. Cryptocurrency and Energy Market Interactions. *J. Financ. Data Sci*. **2020**, *6*, 1–14.
29. Shrimali, G. Market Integration and Renewable Energy Policy. *Renew. Energy* **2019**, *130*, 1150–1161.
30. Gargallo, P.; Salvador, M.; Ventura, S. Energy Portfolios and Risk Management. *J. Risk Finance* **2022**, *23* (2), 178–199.
31. Gujarati, D. N. *Basic Econometrics*, 4th ed.; McGraw-Hill: New York, 2002.

Author

Jeremy Cheung is a rising senior studying at Reading School (Reading, England). He plans to major in economics and has various other experiences in economic research. He is passionate about using this knowledge to create lasting social impact.

DOI: 10.36838/v7i8.1

IJHSR
ijhighschoolresearch.org
REVIEW ARTICLE

The Role of Epigenetics in Association with Type 1 Diabetes

Kalina G. Mihova

Needham High School, 609 Webster St, Needham, MA, 02494, USA; kalinamihova07@gmail.com

ABSTRACT: Type 1 Diabetes (T1D) is an autoimmune disease that occurs when pancreatic β cells produce little or no insulin, thereby leading to an accumulation of glucose in the bloodstream and inability to process carbohydrates. In recent years, T1D has been increasingly researched due to the increased incidence of T1D. While genetics has been studied as a cause for T1D for decades, there have been new studies researching the role of epigenetics and environmental factors in association with T1D. Epigenetics is the study of gene expression and inheritance in which mechanisms such as DNA methylation, histone methylation, and non-coding RNA are used to silence and express specific genes. As understanding of epigenetics increases, there is a clearer correlation between epigenetics and autoimmune diseases. Despite the increase in knowledge on T1D, there is no cure for the disease, and more research will be required to elucidate the complete etiology of T1D and the role of epigenetics in T1D in its natural history. This systematic review describes the role of epigenetics and the environmental factors of maternal health, enteroviruses, age of food introduction, nitrate consumption, and psychological stress in T1D, and summarizes knowledge from current studies to aid in future research.

KEYWORDS: Biomedical and Health Sciences, Genetics and Molecular Biology, Disease Epigenetics, Type 1 Diabetes.

Introduction

Epigenetics:

Beginning in the 19th century, advances in embryology and cell biology led to an initial, broader definition of epigenetics. Up until the 1950s, epigenetics referred to the process by which the fertilized zygote passed on its traits. This definition later evolved to more specifically describe epigenetics as heritable changes in gene expression that are not expressed as changes in the DNA sequence.[1] Since "epi" refers to the above, the epigenome refers to the epigenetic marks layered over cells. These marks or epigenetic mechanisms are able to control whether a gene is expressed or silenced.[2]

Type 1 Diabetes:

Type 1 Diabetes (T1D) is an autoimmune disease in which the pancreatic β cells that produce insulin are destroyed (Figure 1). This destruction leads to hyperglycemia, a high concentration of glucose in the blood, since insulin facilitates the movement of glucose into the cells and stabilizes blood glucose levels.[3] This heightened concentration of glucose in the bloodstream causes weight loss, severe organ damage, heart failure, and early death.[4] In 1921, Frederick Banting discovered insulin and later made it available to the public. This discovery is still used as the primary treatment for T1D and has allowed diabetics to lead much longer and healthier lives. Although the exact causes of T1D and a cure for T1D have not yet been discovered, there have been numerous new technological advancements since the discovery of insulin to aid with insulin delivery.[5] Advancing clinical research and creating new treatments have gained increased focus as there has been a global increase in patients with T1D of about 3-4% in only the last 30 years, likely due to a combination of genetic and environmental factors.[6]

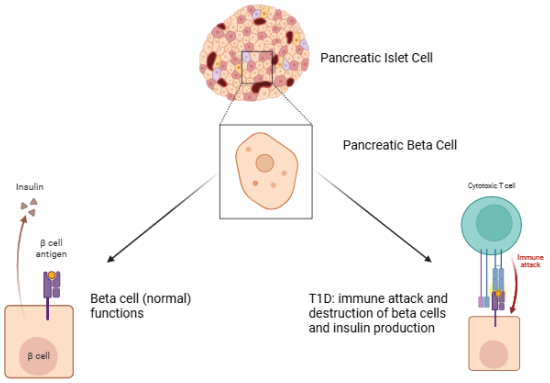


Figure 1: Figure 1 illustrates that pancreatic β cells are part of the pancreatic islet cells. On the left, there is a β cell of an individual without T1D. This cell can carry out regular functions, such as producing insulin and β-cell antigens, without any disturbances. In contrast, the right-hand side displays how a normal β cell may experience an immune attack in a person developing T1D. This immune attack then hinders the cell from carrying out regular functions, such as insulin production, and ultimately leads to an overall reduction in the number of β cells. (Created in BioRender.com)

New Diabetes Treatments:

T1D is commonly treated using continuous blood glucose monitors to measure blood sugar and insulin pumps or injections to stabilize blood glucose levels. These treatments are constantly being researched and improved upon to facilitate T1D management. For instance, one of the recently created treatments for T1D can delay the onset of diabetes by a couple of years. It does so by binding to an epitope of the CD3-epsilon chain expressed on mature T lymphocytes to manage the onset of immune system responses that occur in the development of the disease.[7] Another example of a new treatment is islet cell

DOI: 10.36838/v7i8.2

replacement therapy. In pursuing this therapy, embryonic stem cells offer a functional cure for patients of the disease by acting as pancreatic β cells and releasing insulin to regulate blood glucose levels. However, these replacement therapy options still face significant development challenges, and thus remain largely unavailable to patients and with relatively high patient risks and low adoption, where available, often in limited clinical or research settings.[8] As a result, there is an unmet need for treatments and cures for T1D.

Epigenetics in T1D:

Over the past 40 years, there has been mounting evidence that T1D can be caused by environmental factors.[5] There are over 60 susceptible genes for T1D, most of which are within the HLA region.[9] However, epigenetic changes in the immune system have been discovered in children prior to their development of T1D.[10] This means that some developments of T1D are likely caused by a prior disease and/or external environmental factors that lead to immune system disruption. The immune system then fails to attack accurately and instead damages the pancreatic β cells that produce insulin. Since disease, an external environmental factor, is hypothesized to lead to T1D, epigenetics may be involved.[11] That is because environmental factors, such as disease, use epigenetic mechanisms to express or suppress the underlying genes that factor into the islet cells' production of insulin. By looking at the current understanding of the connections between epigenetics and T1D, as well as recent treatments, we can more clearly understand the specific causes of T1D and use this information to inform ongoing research efforts.

Epigenetic mechanisms:

DNA methylation:

DNA methylation is an epigenetic mechanism involving the formation of a heritable mark created by a methyl group's covalent bond to the C5 position in the cytosine rings of DNA. The heritable mark caused by DNA methylation serves as a method of epigenetic silencing of transcription.[12]

Histone modifications:

Histone modifications regulate chromatin and transcription without expressing any change to the DNA sequence.[13] Histones are groups of proteins found in chromatin that act as a structure on which the DNA strands are tightly coiled. The expression of genes is repressed when the genes are bound by the histones and expressed when the DNA unwinds. This winding or unwinding occurs when epigenetic modifications change the charge of the histones and the DNA. Since the opposite charges of the histones and DNA keep the DNA wrapped (histones have a positive charge while DNA has a negative charge), changing the charge of one or both would also change the gene expression. There are nine different types of histone modification: acetylation, methylation, phosphorylation, and ubiquitylation are the most well-known modifications.[14]

Histone acetylation:

Histone acetylation occurs because histones are positively charged due to lysine and arginine. When acetylation neutralizes the lysine, the histones lose their positive charge and unwind from the DNA. As a result, histone acetylation increases gene expression because it allows for the expression of previously bound genes. This process of acetylation is catalyzed by histone acetyltransferases.[15]

Histone methylation and demethylation:

Histone methylation is the addition of methyl groups onto the histone protein. Histone methylation often prevents transcription factors from binding because it tightens histone tails around DNA, which decreases transcription. Histone demethylation, on the other hand, loosens the tails and leads to gene expression and an increase in transcription. This means that histone methylation can both activate and repress genes.[15]

Non-coding RNA:

Non-coding RNA (ncRNA) are functional RNA molecules that are not translated into protein. ncRNA regulates epigenetic mechanisms and controls the enzymes that catalyze processes such as DNA methylation, histone methylation and demethylation, and histone acetylation. NcRNA is important for understanding many autoimmune diseases because of its role in cell differentiation and tissue development.[16] NcRNA is also important to understanding autoimmune diseases because it is a component that leads to the immune system's attacking healthy tissue, which connects to how diseases attacking the pancreas are a potential cause of T1D.[17]

This up-to-date literature review will discuss the involvement of epigenetics in T1D and aim to provide a topic of discussion for future applications of epigenetics. Specifically, epigenetic and environmental factors such as the age of food introduction, consumption of nitrate, psychological stress, infancy, maternal health, and enteroviruses will be explored for their potential role in the development of T1D.

■ Discussion

Evidence of involvement of T1D with epigenetics:

Epigenetics is pointed out as a potential contributor to the development of T1D because evidence points to T1D not being attributed to genetic causes. Recent studies use the high discordance rate of monozygotic twins, known as MZ twins, in their development of T1D to prove the involvement of environmental factors. Since MZ twins come from the same fertilized egg and share 100% of their DNA, there must be environmental factors that are not part of the DNA sequencing that lead to only one twin developing the disease. Furthermore, the differences in incident rates of T1D between different countries also point to the involvement of environmental factors because there is no other explanation for such a rapid increase or such drastic differences. For instance, sub-Saharan Africa has an incidence rate of only 6% while parts of Scandinavia have a 77% incidence rate.[18] While many studies agree that this high percentage must be due to a common environmental trigger, this environmental trigger remains unclear.

Some possible environmental factors that could explain this difference in incidence rates are explored in the discussion and can be categorized as taking effect either before birth, during infancy, or post-infancy (Figure 2). While other environmental factors have been linked to the development of T1D, the diagram below represents a sample of environmental factors in each phase of development that have been identified with significant odds ratios in other literature.

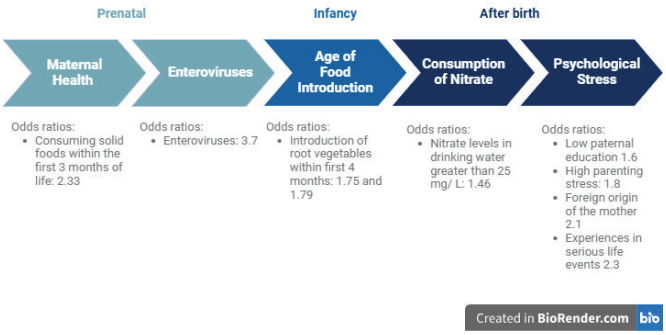


Figure 2: Figure 2 summarizes the main environmental factors associated with T1D included in the discussion. The figure categorizes the factors based on the stages of human development: prenatal, during infancy, or after birth. Odds ratios are included beneath each factor. (Created in BioRender.com)

Epigenetic and environmental factors associated with T1D:

Age of food introduction:

The development of T1D is potentially correlated with the age of food introduction, especially the introduction of root vegetables from an early age. For instance, the development of T1D, genetically, would be more likely with a certain human leukocyte antigen (HLA) system type since the HLA region is the region that is seen as predisposing a patient for T1D (there are 60 such regions). However, among siblings with different HLA types, the age of development is a more prominent factor than the HLA region type and predisposition.[18] Furthermore, studies have pointed to the early introduction of root vegetables, wheat, rye, oats, barley cereals, and eggs in the diet as potential risk factors for the development of T1D. One study has found that the early introduction of root vegetables in a child's diet, meaning in the first four months of life, leads to an increased risk of ß-cell autoimmunity among Finnish children. Children in the two groups that introduced root vegetables within the first four months experienced odds ratios of 1.75 and 1.79, where the baseline is 1. Therefore, the early introduction of root vegetables could be a contributor to the development of T1D.[19]

Consumption of nitrate:

Studies have shown a link between the consumption of toxins in food and water and the development of T1D. Specifically, nitrate and nitrosamine contamination in water is associated with an increase in T1D cases.[20] There have been conflicting studies in the Netherlands demonstrating that the correlation is only present for nitrate levels > 25 mg/ L.[21] When the nitrate levels in drinking water are greater than 25 mg/ L, there is an increased incidence ratio of 1.46, where the baseline level is 1. However, the study had a small sample size since only 15 out of the 1,064 were examined for this factor, which must be taken into account when judging the statistical significance of the data. Furthermore, a stronger correlation between drinking water with nitrate and the development of T1D was demonstrated in a study conducted in northern England. This study demonstrated a significant increase in patients as the nitrate in the water increased.[22]

Psychological stress:

Stress on β-cells is proven to negatively affect the immune response and lead to insulin resistance or the development of T1D. Factors such as rapid growth, trauma, and serious life events, such as the death of a family member, are all shown to increase the demand on β-cells.[23] An ABIS study with a baseline level of 1 demonstrates how physiological factors increase the odds ratios. For instance, high parenting stress leads to an odds ratio of 1.8, experiences in serious life events 2.3, foreign origin of the mother 2.1, and low paternal education 1.6.[24] This β-cell demand leads to non-functioning protein synthesis, proinsulin peptide degradation, and hybrid insulin peptide synthesis through transpeptidation, eventually triggering islet autoimmunity.[23]

Infancy and maternal health:

The development of T1D is potentially correlated with the early introduction of solid foods and the consumption of milk after infancy. A study examining the effects of breastfeeding on children of 3-12 months and the appearance of 4 types of islet antibodies found that an early introduction of solid food is associated with a higher risk of islet autoimmunity for children up to 3 years of age. For instance, the odds ratio for the development of T1D when consuming solid foods within the first 3 months of life is 2.33, where the baseline is 1.[25] Furthermore, a trial on the effects of frequent cow milk consumption found that elevated cow milk antibody concentrations and increased consumption of milk after infancy lead to a higher likelihood of developing islet autoimmunity.[26] In contrast, there has been no association discovered between maternal consumption of gluten, iron, and vitamin C and the risk of the child developing T1D.[27, 28] While maternal consumption of potential environmental factors does not highly impact the development of T1D, the consumption of solid foods during infancy and the consumption of milk after infancy are potential factors for the development of T1D.

Enteroviruses:

Enteroviruses have been researched for involvement in the destruction of pancreatic β cells and the development of T1D due to their ability to suppress immune responses. Enteroviruses are a group of single-stranded RNA viruses, such as the coxsackievirus. One study found that pancreatic β cells are prone to enterovirus infections that often destroy the cells. For instance, the Coxsackievirus B virus family was associated with T1D development due to its impact on the immune system.[29] A study measuring viral protein in the blood of pre-diabetes and diabetes patients found an odds ratio of 3.7, where the baseline is 1. This significant increase in the likelihood of developing T1D when an enterovirus infection is present means

DOI: 10.36838/v7i8.2

that there is an association between enteroviruses and the development of T1D.[30]

Current studies:

Epigenetics and its role in the development of T1D are increasingly being researched to discover causes for T1D. For instance, a study by Johnson *et al*. found 10 regions in the genome that differ between patients with T1D and people without the disease.[31] Furthermore, one study discovered that the methylation at the sites that are associated with the development of T1D can be detected years before the patient begins to develop the disease. This study identified specific autoantibodies and 88 CpG methylation sites that exhibit a correlation with the future onset of T1D, and therefore warrants further investigation.[32] As a result, it is now possible to use these 10 regions and specific site findings to screen patients who are at high risk of developing T1D to monitor or prevent these patients from developing the disease.

Conclusion

There is a rapid increase in patients with T1D, and novel treatments to delay the onset of T1D have been created. However, no permanent cure exists yet, and the exact causes for the disease remain unknown. Therefore, insight into potential epigenetic causes for T1D could aid in future research on the causes of T1D. It is hypothesized that environmental factors are associated with the development of T1D due to the high discordance rate of MZ twins and the genetically inexplicable increase in patients with T1D. Specifically, some environmental factors that could be contributors to the development of T1D are the age of food introduction, the consumption of nitrate, psychological stress, maternal health, and enteroviruses.

Future studies:

In the future, screening high-risk patients for pre-diabetes and researching environmental factors and epigenetic contributors can help lead to a better understanding of the complex etiology of T1D, which involves potentially heritable genetic factors, epigenetic conditions, and multiple intertwined epigenetic factors. Many other factors exist, such as cold climate, vitamin D deficiency, pollution and heavy metal exposure, toxin exposure, childhood obesity, and the gut microbiome, that should be considered in the future. In addition, while this study looks at these factors in isolation, it provides an important jumping-off point for these factors to be studied concurrently in real-world settings, which will help improve understanding of each factor's contribution. Furthermore, since T1D is associated with DNA methylation and histone modification, the creation of epigenetic treatments for T1D is plausible. For instance, histone modification inhibitors function as epigenetic treatments for cancer and therefore have great potential as future therapeutic, or even curative, interventions for T1D that merit future research.

Acknowledgments

This paper was written with the guidance of a mentor through the LearnSTEM program.

References

1. Jerram ST, Dang MN, Leslie RD. The Role of Epigenetics in Type 1 Diabetes. Curr Diab Rep. 2017;17(10):89.
2. Riggs AD. X inactivation, differentiation, and DNA methylation. Cytogenet Cell Genet. 1975;14(1):9-25.
3. DiMeglio LA, Evans-Molina C, Oram RA. Type 1 diabetes. Lancet. 2018;391(10138):2449-62.
4. Siang Yong Tan JM. Frederick Banting (1891-1941): Discoverer of Insulin. Singapore Medical Journal. January 2017.
5. Atkinson MA. The pathogenesis and natural history of type 1 diabetes. Cold Spring Harb Perspect Med. 2012;2(11).
6. Norris JM, Johnson RK, Stene LC. Type 1 diabetes-early life origins and changing epidemiology. Lancet Diabetes Endocrinol. 2020;8(3):226-38.
7. Vertex Announces Positive Results From Ongoing Phase 1/2 Study of VX-880 for the Treatment of Type 1 Diabetes Presented at the American Diabetes Association 84th Scientific Sessions [press release]. Vertex Pharmaceuticals IncJune 21, 2024.
8. Ramos EL, Dayan CM, Chatenoud L, Sumnik Z, Simmons KM, Szypowska A, *et al*. Teplizumab and beta-Cell Function in Newly Diagnosed Type 1 Diabetes. N Engl J Med. 2023;389(23):2151-61.
9. Xie Z, Chang C, Huang G, Zhou Z. The Role of Epigenetics in Type 1 Diabetes. Adv Exp Med Biol. 2020;1253:223-57.
10. Zhang J, Chen LM, Zou Y, Zhang S, Xiong F, Wang CY. Implication of epigenetic factors in the pathogenesis of type 1 diabetes. Chin Med J (Engl). 2021;134(9):1031-42.
11. Al Aboud NM, Tupper C, Jialal I. Genetics, Epigenetic Mechanism. StatPearls. Treasure Island (FL)2024.
12. Jin B, Li Y, Robertson KD. DNA methylation: superior or subordinate in the epigenetic hierarchy? Genes Cancer. 2011;2(6):607-17.
13. Alaskhar Alhamwe B, Khalaila R, Wolf J, von Bulow V, Harb H, Alhamdan F, *et al*. Histone modifications and their role in epigenetics of atopy and allergic diseases. Allergy Asthma Clin Immunol. 2018;14:39.
14. Greer EL, Shi Y. Histone methylation: a dynamic mark in health, disease and inheritance. Nat Rev Genet. 2012;13(5):343-57.
15. Lee HT, Oh S, Ro DH, Yoo H, Kwon YW. The Key Role of DNA Methylation and Histone Acetylation in Epigenetics of Atherosclerosis. J Lipid Atheroscler. 2020;9(3):419-34.
16. Pathania AS. Crosstalk between Noncoding RNAs and the Epigenetics Machinery in Pediatric Tumors and Their Microenvironment. Cancers (Basel). 2023;15(10).
17. Kumar D, Sahoo SS, Chauss D, Kazemian M, Afzali B. Non-coding RNAs in immunoregulation and autoimmunity: Technological advances and critical limitations. J Autoimmun. 2023;134:102982.
18. Akil AA, Jerman LF, Yassin E, Padmajeya SS, Al-Kurbi A, Fakhro KA. Reading between the (Genetic) Lines: How Epigenetics is Unlocking Novel Therapies for Type 1 Diabetes. Cells. 2020;9(11).
19. Virtanen SM, Takkinen HM, Nevalainen J, Kronberg-Kippila C, Salmenhaara M, Uusitalo L, *et al*. Early introduction of root vegetables in infancy associated with advanced ss-cell autoimmunity in young children with human leukocyte antigen-conferred susceptibility to Type 1 diabetes. Diabet Med. 2011;28(8):965-71.
20. Moltchanova E, Rytkonen M, Kousa A, Taskinen O, Tuomilehto J, Karvonen M, *et al*. Zinc and nitrate in the ground water and the incidence of Type 1 diabetes in Finland. Diabet Med. 2004;21(3):256-61.
21. van Maanen JM, Albering HJ, de Kok TM, van Breda SG, Curfs DM, Vermeer IT, *et al*. Does the risk of childhood diabetes mellitus require revision of the guideline values for nitrate in drinking water? Environ Health Perspect. 2000;108(5):457-61.
22. Parslow RC, McKinney PA, Law GR, Staines A, Williams R, Bodansky HJ. Incidence of childhood diabetes mellitus in York-

shire, northern England, is associated with nitrate in drinking water: an ecological analysis. Diabetologia. 1997;40(5):550-6.
23. Rewers M, Ludvigsson J. Environmental risk factors for type 1 diabetes. Lancet. 2016;387(10035):2340-8.
24. Sepa A, Wahlberg J, Vaarala O, Frodi A, Ludvigsson J. Psychological stress may induce diabetes-related autoimmunity in infancy. Diabetes Care. 2005;28(2):290-5.
25. Hakola L, Takkinen HM, Niinisto S, Ahonen S, Nevalainen J, Veijola R, *et al*. Infant Feeding in Relation to the Risk of Advanced Islet Autoimmunity and Type 1 Diabetes in Children With Increased Genetic Susceptibility: A Cohort Study. Am J Epidemiol. 2018;187(1):34-44.
26. Niinisto S, Cuthbertson D, Miettinen ME, Hakola L, Nucci A, Korhonen TE, *et al*. High Concentrations of Immunoglobulin G Against Cow Milk Proteins and Frequency of Cow Milk Consumption Are Associated With the Development of Islet Autoimmunity and Type 1 Diabetes-The Trial to Reduce Insulin-dependent Diabetes Mellitus (IDDM) in the Genetically at Risk (TRIGR) Study. J Nutr. 2024;154(8):2493-500.
27. Mattila M, Hakola L, Niinisto S, Tapanainen H, Takkinen HM, Ahonen S, *et al*. Maternal Vitamin C and Iron Intake during Pregnancy and the Risk of Islet Autoimmunity and Type 1 Diabetes in Children: A Birth Cohort Study. Nutrients. 2021;13(3).
28. Hakola L, Lund-Blix NA, Takkinen HM, Tapanainen H, Niinisto S, Korhonen TE, *et al*. Maternal gluten, cereal, and dietary fiber intake during pregnancy and lactation and the risk of islet autoimmunity and type 1 diabetes in the child. Clin Nutr ESPEN. 2024;62:22-7.
29. Sarah J Richardson NGM. Enteroviral infections in the pathogenesis of type 1 diabetes: new insights for therapeutic intervention. Curr Opin Pharmacol. December 2018.
30. Yeung WC, Rawlinson WD, Craig ME. Enterovirus infection and type 1 diabetes mellitus: systematic review and meta-analysis of observational molecular studies. BMJ. 2011;342:d35.
31. Johnson RK, Vanderlinden LA, Dong F, Carry PM, Seifert J, Waugh K, *et al*. Longitudinal DNA methylation differences precede type 1 diabetes. Sci Rep. 2020;10(1):3721.
32. Stefan M, Zhang W, Concepcion E, Yi Z, Tomer Y. DNA methylation profiles in type 1 diabetes twins point to strong epigenetic effects on etiology. J Autoimmun. 2014;50:33-7.

Author

Kalina Mihova is a junior at Needham High School, MA. Kalina has been living with T1D since 2017 and is passionate about clinical research. Kalina is a Level 9 rhythmic gymnast and is involved with piano and debate. She is interested in pursuing biology or genetics at a leading university.

DOI: 10.36838/v7i8.2

IJHSR
ijhighschoolresearch.org
REVIEW ARTICLE

Cellular Origin of Glioblastoma: Current Evidence, Challenges, and Implications

Grace R. Qian[1], Yiwei Fu[1], Albert H. Kim[1, 2 *]

1) Taylor Family Department of Neurosurgery, Washington University School of Medicine, St. Louis, MO, 63110, USA
2) The Brain Tumor Center, Siteman Cancer Center, Washington University School of Medicine, St. Louis, MO, 63110, USA; alberthkim@wustl.edu

ABSTRACT: Glioblastoma (GBM) is the most common primary cancer of the central nervous system (CNS) in adults, with a 5-year relative survival rate of 6.9%. For the past 30 years, standard treatment has included a combination of surgical resection with radiation therapy and temozolomide. Progress in treatment for GBM has been hindered due to the brain's limited repair abilities, GBM's diffuse nature into eloquent brain areas that make full resection essentially impossible, and the heterogeneous tumor, which contributes to treatment resistance and inevitable recurrence. As single-cell RNA sequencing allows for identifying subtumoral cellular populations, a potential research area is to study the cell of origin—the cells that accumulate specific mutations in the right conditions to become tumorigenic. Studying the basic science behind transforming the cell of origin into GBM offers insight into how GBM may recur and develop targeted drugs. Here, we provide a comprehensive literature review on possible cells of origin, including neural stem cells (NSCs), oligodendrocyte progenitor cells (OPCs), and astrocytes. We conclude that the cellular origin of GBM in addition to specific mutations and environmental conditions, may better define a specific patient's GBM, which has implications for diagnosis, therapy, and prognosis.

KEYWORDS: Biomedical and Health Sciences, Genetics and Molecular Biology of Disease, Glioblastoma.

Introduction

Glioblastoma (GBM) is a grade IV highly aggressive primary malignant glioma, which accounts for 48.6% of malignant central nervous system tumors, making it the most common primary cancer of the central nervous system (CNS) in adults.[1,2] The median age of diagnosis is 64 years, and the incidence increases with age, peaking at 75–84 years.[3] The median survival is 15 months post-diagnosis with a 5-year survival rate of 6.9%. Among individuals with GBM, prominent figures including Beau Biden and senators Ted Kennedy and John McCain have been afflicted with this disease.[4]

Multiple challenges arise in treating recurrent GBM, including the brain's limited regenerative capacities, the unique and selectively permeable blood-brain barrier (BBB) vascularization that makes drug delivery difficult, GBM's high invasiveness and infiltration into eloquent brain areas that renders full resection essentially impossible, and tumor resistance to radiation and chemotherapy — which all lead to inevitable recurrence.[2,5] GBMs have been classified into three transcriptional subtypes – Proneural, Classical, and Mesenchymal.[6] Currently, standard treatment includes surgical resection to debulk the tumor, followed by fractionated radiation therapy and chemotherapy with temozolomide.[7,8] Despite intensive treatment, GBM has a near 100% recurrence rate, with a 10-year survival rate of only around 1%.[4]

There has been renewed interest in the "cell of origin" concept in recent years. Exact definitions of the cell of origin vary, but consensus agrees that the normal cell is malignantly transformed into the first GBM cell.[9] There are multiple possible cells of origin for GBM. The cell of origin, combined with specific mutations, may determine the specific molecular features of an individual's GBM, thus offering a potentially useful means of stratifying GBM tumors for distinct therapeutic strategies. Identifying the cell of origin of GBM could also help us understand the biological mechanisms of tumor initiation and provide additional and possibly earlier targets in the process of GBM pathogenesis. Here, we discuss evidence for neural stem cells (NSCs), oligodendrocyte progenitor cells (OPCs), and astrocytes for being key cell types of origin in GBM development. The mechanism of transformation is debated, with theories of accumulating somatic genetic mutations, dedifferentiation of progeny cells, and epigenetic changes accounting for recurrence. Importantly, molecularly distinct tumors formed when the same mutations—NF1, Trp53, and PTEN—were mutated in NSCs and OPCs, suggesting that the cell of origin has implications for tumor phenotype.[10] The mutational signature distinctions that arise from cell-of-origin differences may enable the characterization of properties and therapeutic vulnerabilities. Importantly, there is the distinction between cells of origin and cells of mutation; The cell of mutation is the cell in which the DNA mutation occurs, potentially because of DNA damage. On the other hand, the cell of origin is the cell in which the mutation itself is manifested biologically and acquires malignant features. For example, in familial cancer, mutations may be harbored in multiple cell types (cells of mutation), but only certain cells progress into tumors (cells of origin). Distinguishing between cells that transform into malignant tumors and those that aquire initial mutations is important since identifying the cell of origin may provide insight into tumor development, GBM subtypes, and potential future

DOI: 10.36838/v7i8.3

therapies. Thus, identifying and considering the cell of origin in combination with the specific mutations and environmental conditions may improve personalized and patient-specific treatment in the future.

Discussion

In understanding the cellular origins of GBM, it is important to introduce normal glial development (Figure 1). Neural stem cells in specialized niches give rise to glial and neural progenitor cells, which differentiate into oligodendrocyte and astrocyte lineages. There is support for these progenitor pools as potential cells of origin. This evidence is summarized in Table 1.

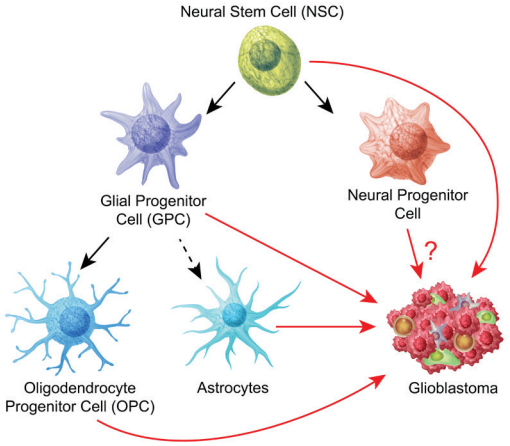


Figure 1: Overview of developmental lineage and GBM cells of origin, highlighting the developmental relationship between possible cells of origin. Neural stem cells (NSCs) give rise to neural progenitor cells and glial progenitor cells (GPCs), which in turn differentiate into oligodendrocyte progenitor cells (OPCs) and astrocytes. These cell types are all able to accumulate mutations and form GBM.

Table 1: Summary and key findings of experimental evidence for specific cell of origin-mutation combinations and environmental conditions. NSCs, OPCs, and astrocytes were all able to successfully produce GBM tumors in various mouse models and cell lines.

Cell of Origin	Driver Mutation	Key findings/ important notes	Reference
NSCs	Nf1/p53 or Nf1/p53/Pten	Tamoxifen-inducible *nestin-creER*^{T2} in vivo mice model	Alcantara Llaguno et al. 2009 [18]
	Trp53/Pten/Egfr	Mouse model, only SVZ injections produced tumors	Lee et al. 2018 [13]
	Rb/p53, Rb/p53/Pten, or p53/Pten	Mouse, mature astrocytes did not produce tumors	Jacques et al. 2010 [21]
	Pten	Mesenchymal stem cells did not produce tumors	Duan et al. 2015 [19]
OPCs	Pdgfa	Retroviral vector, tumors expressed OPC markers (NG2, PDGFRA)	Assanah et al. 2006 [33]
	Pdgfb	PDGF is the mitogen for OPCs	Lindberg et al. 2009 [30]
	S100b promoter Egfr/p53	S100b promoter is not expressed in NSCs	Persson et al. 2010 [67]
	p53/Nf1	Mosaic Analysis with Double Markers (MADM), no tumor production in NSC lines	Liu et al. 2011 [47], Zong et al. 2005 [48]
Astrocytes	P16INK4a/p19ARF	SCID mouse, both NSCs and astrocytes formed GBM	Bachoo et al. 2002 [38]
	p53/Pten/Rb	Fully penetrant mouse model, >20% tumors formed in non-neural progenitor niches	Chow et al. 2011 [39]
	Reinduction of Erbb2	Mouse, formed radial glial progenitors (dedifferentiation)	Ghashghaei et al. 2007 [35]

Abbreviations: Subventricular zone (SVZ), Platelet-derived growth factor (PDGF) – A/B denotes subunits, Neurofibromatosis Type 1 (NF1), Phosphatase and tensin homolog deleted on chromosome 10 (PTEN), Epidermal Growth Factor Receptor (EGFR), Human Epidermal Growth Factor Receptor 2 (HER2/ERBB2), Retinoblastoma (Rb), **Tumor suppressor genes:** PTEN, p53, NF1, Rb, **Tumor oncogenes:** PDGF, EGFR, ERBB2 **Activating mutations:** PTEN, p53, EGFR, ERBB2 **Repressing mutation:** NF1, Rb, PDGF

Neural Stem Cells as a Cell of Origin:

NSCs are a fundamental cell type in the development of the brain, as they give rise to all of the cells of the CNS.[13] NSCs divide both asymmetrically, generating differentiated cell types like neurons, astrocytes, and oligodendrocytes, and progenitor cells like OPCs or GPCs, and symmetrically, generating more NSCs (Figure 1). Their replication and growth contribute to cortical expansion during development. Historically, all neurons were believed to be generated during development before adulthood.[14] However, pioneering work by Nottebohm using labeled DNA precursors found evidence for producing new neurons in adulthood in avian models,[15] suggesting that neurogenesis also occurs later in life through the maintenance and division of a pool of NSCs that persists into adulthood. These postnatal NSCs are localized to the astrocytic ribbon of the brain's subventricular zone (SVZ), adjacent to the lateral ventricles,[16] and the hippocampus's subgranular zone (SGZ). Growing evidence suggests that NSC localization in the SVZ is significant in creating a seed-to-soil relationship for gliomagenesis. The SVZ microenvironment may be a neurogenic niche for NSCs through the release of chemoattractants like pleiotrophin.[17] The SVZ's interactions with cerebrospinal fluid (CSF) and vascularization also provide a rich pro-tumor microenvironment.[18,19] Signals to the SVZ from the CSF and blood are helpful during development in transferring growth factors to proliferate NSCs. Still, disruption of the control system can corrupt tumor formation and growth.[18] Furthermore, in an MRI-based study, 93% of GBMs contacted at least some part of the lateral ventricular wall lined by the SVZ.[20] Thus, the location of GBM in the SVZ offers a unique look into NSCs as a potential cell of origin for GBM. These studies suggest that NPCs are important for glioma invasion by releasing factors like pleiotrophin, which help create an attractive and supportive niche for glioma cells, with further implications that these NSCs themselves may transition into tumor cells.

- NSC Evidence:

Several studies have investigated the tumorigenic potential of neural stem cells harboring oncogenic mutations in mouse models.[21] Deletion of p53, NF1, and PTEN, specifically in embryonic or adult NSCs by the Cre/loxP system, was sufficient to generate GBMs in mice.[22] In addition, conditional knockout of p53, NF1, and PTEN in either adult mouse SVZ, where NSCs reside, or nonneurogenic areas, such as the cortex, using transgenic mice led to tumor formation only in the SVZ.[23,24] Together, these studies suggest that NSCs residing in the SVZ can act as the cells of origin for GBM. Also, it was found that low levels of GBM driver mutations can be detected in SVZ cells, suggesting that tumor cells had originated and migrated

from the SVZ. For directional validation, the authors discovered that while tumor cells exhibited tumor-unique mutations, no SVZ cells contained SVZ-unique mutations, suggesting that the directionality of gliomagenesis was SVZ to the tumor.[25] Upon further analysis, these putative origin cells were NSCs in the astrocytic ribbon. This was confirmed in a mouse model in which p53/PTEN/EGFR mutations were induced in the SVZ, producing tumors in NSC mutant mice; other groups have found similar results.[25,26] Additionally, evidence from human and mouse data suggest that GFAP-positive NSCs may act as the cell of origin. By analyzing TERT promoter mutation enrichment in various SVZ layers, one study found that these mutations were significantly enriched only in GFAP-positive NSCs in the astrocytic ribbon, not in bulk astrocytes. This suggests that GFAP-positive NSCs, rather than mature differentiated astrocytes, may act as the primary cells of origin for GBM.[16]

Furthermore, mutations in the telomerase reverse transcriptase promoter (TERTp) associated with GBM were found in the astrocytic ribbon, where many NSCs reside in the adult brain.[25] TERTp mutations are found in more than 80% of GBMs, suggesting that TERTp may be a critical mutation of GBM. Its presence in the NSC niche further supports NSCs as the cell of origin.[2] Another group showed that human induced pluripotent stem cell lines, upon differentiation into NSCs and deletion of canonical GBM mutations such as PTEN, NF1, TP53, and PDGFRA, could form GBM of the various molecular subtypes when implanted into the brains of immunocompromised mice.[24] CD133 could be a potentially important marker that can help identify both neural stem cells (NSCs) and GBM stem cells. Importantly, however, sharing the same markers does not necessarily imply a lineage relationship between NSCs as a cell of origin and GBM stem cells, a topic that requires further investigation.[27,28] Thus, the ability to malignantly transform human NSCs into tumorigenic GBM cells offers additional evidence that NSCs are a cell of origin for GBM. Secondarily, neurogenic niches that house neural progenitor cells (NPCs) may promote GBM pathogenesis or malignant behaviors. However, to our knowledge, there is no experimental evidence for NPCs per se as cells of origin.

Cells of Origin from Glial Lineage:

Neural stem cells give rise to progenitor cells, which can differentiate into glial cells (Figure 1). Glial cells are the supportive "glue" of the brain and spinal cord, with a ratio of one glial cell to one neuron.[29] Glial cells include astrocytes and oligodendrocyte lineage cells.[30] Their functions involve directing neuronal migration, synaptogenesis, influencing growth, and monitoring the CNS microenvironment.[31] A long history of GBM literature suggests that glial lineage cells may be the cells of origin. Therefore, whether other developmental neural cell states harbor tumorigenic potential remains to be seen.

Oligodendrocyte Progenitor Cells as a Cell of Origin:

Oligodendrocyte progenitor cells (OPCs) are lineage-restricted progenitor cells that arise from the asymmetrical division of NSCs during development and reside in the parenchyma of adult brains.[32] OPCs are more abundant than NSCs and more widely distributed, constituting 70% of the dividing cells in the brain.[9,32,33] They can also arise from the differentiation of glial progenitor cells (GPCs) and constitute one of the specific populations of progenitor cells within the heterogeneous glial progenitor population. OPC-specific markers include neural/glial antigen 2 (NG2), a type of cell membrane glycoprotein or proteoglycan, the PDGF receptor alpha (PDGFRA), and Olig1.[34] Histopathologic analysis of human GBM shows expression of these markers,[10] suggesting that GBM may derive from OPCs as a cell of origin. However, NSCs are progenitor cells for most of the population of cells within the brain. Thus, we could expect that NSCs with tumorigenic potential may go down a developmental route, leading them to express markers found in OPCs instead of the tumor arising in OPCs themselves. Thus, a method to confirm and trace lineage would be important in supporting the OPC theory for the cell-of-origin.

In GBM, PDGFRA mutations are the second most common tyrosine kinase receptor mutation, observed in around 30% of patients.[35] Exogenous PDGFA infusion into the adult SVZ has been shown to induce OPCs to form GBM lesions.[36] In mice, PDGFB transfer via a Ctv-a transgenic mouse model induced gliomas in 33% of the cases.[37] Furthermore, other growth factors, such as epidermal growth factor receptor (EGFR) overexpression, are common in about 60% of primary glioblastomas.[38] The overexpression of growth factors like EGFR and PDGF is catalyzed by mutations in tumor suppressor genes like p53 and PTEN, thus suggesting that p53 and PTEN mutations cannot generate GBM independently but work in coordination with growth factor mutations.[34] These observations have been confirmed in murine models, showing successful tumor formation.[39] Injection of a PDGF growth factor-expressing retrovirus with GFP radiolabeling into the subcortical white matter formed GBM tumors in 100% of animals (86/86). At the same time, none of the GFP-only control mice showed any signs of tumor growth. The replication incompetent retrovirus selectively infects the cycling glial progenitors, and their identity was confirmed by immunohistochemistry staining showing the presence of markers like NG2, OLIG2, and CC1, which are associated with OPCs.[40] The successful tumor formation of OPCs in mice provides evidence that they may represent a possible cell of origin. In summary, successful tumor formation with specific growth factors and tumor suppressor mutations in OPCs provides evidence that OPCs can be the cell of origin in these models. In summary, the same mutations in OPCs generate a molecularly distinct GBM from that of NSCs. The difference in tumorigenesis from NSCs and OPCs highlights how the cell of origin in GBM may determine the specific subtype of GBM and may help to stratify GBM tumors for distinct therapeutic strategies.

DOI: 10.36838/v7i8.3

Astrocytes as a cell of origin:

-Astrocyte dedifferentiation theory and mutation dependency:

Astrocytes are a type of glial cell. They are the most abundant glial cells and link neurons to the blood supply, forming a critical blood-brain-barrier (BBB) component.[30] Embryologically, they develop from outer radial glia (oRG) cells and subpallial cells near the basal ganglia.[41] Astrocytes develop from neural progenitors in the SVZ in adults and proliferate through local mitotic division. At the end of embryonic development, oRG cells give rise to astrocytes. Early on, it was shown that mature mouse astrocytes can dedifferentiate to radial glia (which don't exist in adult brains) through *in vivo* induction of the tyrosine kinase receptor ErbB2. When cultured, these dedifferentiated glial progenitors can give rise to mature glial cells like astrocytes. These dedifferentiated astrocytes share similar properties with embryonic oRG cells and are less lineage-restricted than typical adult NSCs.[42]

Additionally, astrocytes cultured with TGFα were converted to radial glial cells and subsequently differentiated into NSCs that formed self-renewing neurospheres.[43] These unique properties of astrocytes offer insight into the astrocyte dedifferentiation theory and lay a foundation for studies of astrocyte gliomagenesis. One study showed that upregulation in growth factor receptors like PDGF contributes to the activation of autocrine loops that lead to high-grade astrocytomas.[44] Similarly, the combined loss of p16INK4a and p19ARF triggered the dedifferentiation of astrocytes in response to EGFR activation, forming GBM-like tumors upon orthotopic implantation into mice.[38] A mouse model with Tp53/PTEN deletion mutations induced in the SVZ targeting GFAP (a marker for astrocytes) positive cells found that over 20% of tumors were formed in regions outside of the SVZ. Thus, it is difficult to distinguish between astrocytes and NSCs in the SVZ. Still, it was hypothesized that the tumors formed in the cortex, brain stem, cerebellum, and spinal cords originated from astrocytes.[45] One study found that sequential mutation of NF1 and p53 in isolated astrocyte cultures had weak transforming abilities. But, when the same mutant astrocytes were treated with PDGF and EGF growth factor, they could form tumors when injected into mouse brains. Thus, astrocytes can transform into tumor cells, but this phenomenon may depend on the mutation and growth factor exposure.[46] Astrocytes with mutations in the retinoblastoma protein family initiated grade II gliomas. However, additional mutations in PTEN caused grade progression, suggesting that gliomagenesis depends not only on the cell of origin but also on the mutation.[9,47] Importantly, modeling with genetically engineered mice (GEMM) is not error-proof, and there are limitations to inducing mutations in specific cell populations. In summary, evidence for the astrocyte dedifferentiation theory and successful GBM formation in mice suggest that astrocytes are a possible cell-of-origin.

- Debate over biomarkers and astrocytes as cells of origin:

Dedifferentiation of astrocytes or other terminal cell populations has been a controversial theory since evidence for this is derived primarily from *in vitro* and *in vivo* mouse studies rather than human studies.[12,48] In particular, in the transgenic mouse experiments, it is challenging to ensure that only terminal astrocyte populations are targeted (*e.g.*, GFAP-driven transgenes are expressed in both astrocytes and NSCs). In studying the cell of origin for GBM, Glial Fibrillary Acid Protein (GFAP) is generally used as a promoter or marker for astrocytes. However, it has been difficult to find markers that account for the heterogeneity of the astrocytic population uniquely in the astrocyte population. This is the case for GFAP expressed by all astrocytes and NSCs.[41] Additionally, GFAP+ cells don't always account for all astrocytes, as more than 40% of astrocytes in mice were GFAP-. [49,50] There have also been difficulties in tracing the lineage of astrocytes through biomarkers, most recently with limited successes in determining molecular markers for the astrocyte precursor cell (APC). In conclusion, the shared biomarkers make it challenging to create transgenic mice that target astrocytes specifically. Furthermore, findings from studies that use a GFAP promoter are thus subject to uncertainties of whether astrocytes or NSCs induced tumorigenesis.

Thus, the non-specificity of transgenic mouse experiments and the difficulty in distinguishing mature astrocytes from NSCs and even earlier neural progenitors using cellular markers are all reasons that astrocytes have remained controversial as a cell of origin for GBM.

However, astrocyte progenitor cells may still have potential as cells of origin. According to Lee *et al*, GBM cells of origin are a GFAP-positive NSC or neural progenitor in the astrocytic ribbon of the SVZ.[16] The non-specificity of GFAP expression and uncertainty of the identity of this cell of origin opens the possibility of a glial progenitor as a potential cell of origin (which can form both OPCs and astrocyte progenitor cells). Thus, while the dedifferentiation theory is controversial and there is no direct evidence for astrocytes as cells of origin *in vivo* and in human tumors, it is possible that astrocytes or other glial progenitors may act as a cell of origin. Future research is needed to elucidate the identity of these cells.

Mesenchymal Stem Cells:

Mesenchymal stem cells (MSCs) are a type of multipotent adult stem cells. They can differentiate into various tissue cell precursors like adipocytes, chondrocytes, and osteoblasts. In GBM, MSCs can arise locally or differentiate from glioma stem cells (GSCs). They exhibit tumor-promoting effects such as suppressing immune responses and supporting tumor growth. Single-cell RNA sequencing has provided evidence for MSCs as a possible cell of origin for the GBM mesenchymal subtype.[51]

Despite this evidence, the study of MSCs as a candidate cell of origin for GBM is still in its infancy. The mesenchymal subtype is also defined by other factors such as the influence of the tumor microenvironment, accumulating mutations like NF1, and treatment-induced mesenchymal transition (wherein treated tumors tend to shift towards the more resistant mesenchymal phenotype.[51] Nevertheless, recent transcriptomic and lineage-tracing studies have begun to reveal the potential of MSCs as cells of origin.

Additional Challenges in Defining Cell of Origin Studies:

- Plasticity:

The plasticity of cell lineages creates another nuance in the research on the cell of origin of GBM by revealing how tumors can evolve between origination, progression, and recurrence. Research on the role of epigenetics (reversible modifications that affect gene expression without altering the DNA sequence) in GBM recurrence has shown that the cell of origin may be more dynamic and fluid and less unchangeable and anchored, as previously thought. For example, it was demonstrated that culturing OPCs with fetal calf serum (FCS), cytokines, and basic fibroblast growth factor (bFGF) could revert OPCs to multipotential NSCs that could differentiate successfully into oligodendrocytes, neurons, and astrocytes.[52] Additionally, OPCs were found to be able to turn into NSCs through reactivation of SOX2 and chromatin remodeling on histone H3.[53] Plasticity thus represent a potential challenge in identifying the cell of origin.

- Cell-autonomous vs. Non-Cell-autonomous:

In transgenic mice models of the cell of origin, mutations cannot be induced in a singular cell but rather in a general population. Thus, these models cannot differentiate between the cell of origin and their neighbors. As such, phenotypic changes cannot be exclusively attributed to mutations in the cell of origin. The effects of a driver mutation depend not only on the mutation itself (cell-autonomous) but also on the genetic background and the tumor microenvironment. These non-cell autonomous effects can affect the cancer cell phenotype, leading to further heterogeneity.[54] Liu *et al*. induced p53/NF1 mutations under the Mosaic Analysis with Double Markers (MADM) system and found that although mutations were initiated in NSCs, only the OPC cell type proliferated and formed GBM tumors.[11] The conclusion was that the cell of mutation (NSC) is distinct from the cell of origin (OPCs). This suggests that OPCs can be the cell of origin, at least with p53/NF1 mutation. The technique of MADM allows for tracing green fluorescent protein-tagged single cells prior to malignant transformation and continuous comparison to a wild-type red fluorescent protein-tagged sibling cell throughout cell cycles and evolution.[55] This addresses the issue of cell-autonomous vs. non-cell-autonomous effects. Thus, additional studies should be conducted with different mutations.[56] The results of these studies would shed light on the relationship between driver mutation and cell of origin in GBM.

- Tumor Heterogeneity:

One major obstacle in developing effective treatments for GBM is its highly heterogeneous nature, both intertumorally and intratumorally. There have been various approaches to studying GBM subtypes. In 2008, the Cancer Genome Atlas created a transcription-based classification system to stratify GBM into four subtypes (Proneural, Neural, Classical, and Mesenchymal), which were also associated with PDGFRA and TP53 mutations for Proneural, EGFR for Classical, and NF1 for Mesenchymal subtypes.[57,58] Verhaak and colleagues found that treatment susceptibility may vary amongst these subtypes.[56] In 2017, Verhakk's group updated their subtyping system, removing the neural subtype after showing it was primarily composed of normal brain elements and thus did not represent an individual subtype.[59]

Despite classification attempts, there are limitations in the current molecular classification system.[60] For example, subtypes can change with recurrence. One study found that 63% of patients were observed with a different transcriptional subtype after recurrence.[61] Furthermore, characterization of the post-treatment proneural-to-mesenchymal transition, or PMT, shows increased aggressiveness, tumor-associated macrophages (TAMs), and therapy resistance.[62,63] Additionally, Single-cell RNA sequencing (scRNA-seq) of tumor cells demonstrated a mixture of subtypes within one tumor, i.e., intratumoral heterogeneity.[64,65] These subpopulations of tumor cells vary in function, type, and lineage. While previous classification based on bulk tumors accounted for intertumoral heterogeneity, developing sensitive single-cell sequencing provides insights into defining a high-resolution classification of GBM subpopulations. Neftel *et al.* used scRNA-seq to identify four distinct tumor cell states, which were also associated with mutations in CDK4, EGFR, PDGFRA, and NF1, respectively.[19,66] Not only does GBM exhibit molecular heterogeneity, but developmental heterogeneity as well. Single-cell techniques have also opened further discussion on the existence of a self-renewing glioma stem cell population (GSC), which may also play a role in GBM recurrence.[65,67] Efforts to more comprehensively characterize the intratumoral heterogeneity of GBMs raise the possibility of stratifying patients based on predicted similarities in response to treatment and targeting specific subcellular populations. For GBM, intratumoral heterogeneity reflects and accounts for intertumoral heterogeneity, so recent focus has been shifted to tumor genetic and transcriptomic classification.[68] In the future, subpopulation-based classification could guide precision therapy.

Therapeutic Implications of the Cell of Origin in GBM:

Currently, standard GBM treatment includes surgical resection followed by radiation and chemotherapy with temozolomide.[69] Clinical trials of Optune, a cancer division prevention device, and the EGFRV3 vaccine have improved survival time.[70,71] However, patients survive only around 17-20 months, even with this novel treatment. Additionally, GBM is well known for its near 100% recurrence rate, which occurs at least in part due to the heterogeneous nature of the tumor and the natural selection that occurs when using drugs like temozolomide that don't target specific tumoral cell populations.[72,73] GBM derived from different cells of origin have shown corresponding differences in drug/treatment sensitivities. Targeting specific cells of origin offers a possible therapy route, such as Dasatinib (an ERBB2 inhibitor), which effectively targets OPC-derived tumors.[58] Furthermore, NSC-originated tumors are less sensitive to temozolomide therapy than OPC-derived tumors.[74] Stratifying glioblastoma in patients may be an important way to determine the most specific treatment possible to improve patient outcomes and prevent recurrence and drug resistance in the progression of the disease (Figure 2).

DOI: 10.36838/v7i8.3

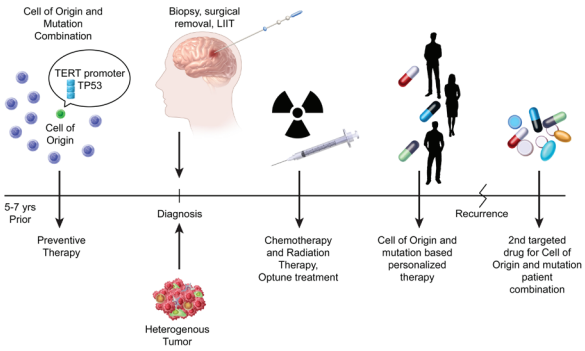


Figure 2: Timeline of standard treatment for GBM and how cell-of-origin-based treatments may be integrated into routine treatment. With a combination of early preventive therapy, personalized therapy, and drugs, GBM treatment may improve patient outcomes and survival.

- Personalized Therapy:

In the future, GBM therapy may target heterogeneous populations with various treatments for each population, including astrocytes, oligodendrocyte progenitors, neural stem cells, glial progenitors, and cancer stem cells. This therapy could be personalized per patient based on the various combinations of cell populations defined by the cell of origin and mutation combinations unique to each patient. Research has shown from patient-derived samples that NSC clustered primary GBM cells were highly malignant and more sensitive to drugs.[75] For example, to target NSC-derived tumors, radiation therapy has been targeted to the SVZ region as a blockade against further gliomagenesis, with significant increases in patient progression-free and total survival in Glioblastoma patients post-total resection.[22] Since GBM tumors tend to be radiation resistant, the use of the chemokine CXCL12, which has been shown to increase the radiosensitivity of the SVZ, offers a possible combination therapy.[76] There have also been suggestions on using DNA technology like CRISPR/Cas9 to correct driver GBM mutations, such as tumor suppressor mutations in mutated NSCs within the SVZ.[18] PTEN deficiency, a common mutation in GBM, has also been shown to allow NSCs to reprogram to a more stem-cell-like state through the indirect effects of increasing PAX7.[24] Thus, upregulating PTEN in the SVZ region offers another possible therapy. Immunotherapy and small-molecule inhibitors can inhibit TERT mutations in the SVZ, which usually produce pro-tumor effects by creating telomerase.[77]

Current clinical trials harness the immune system's ability to specifically target GBM mutations like EGFR. A phase II trial using rindopepimut (a vaccine targeting EGFRvIII) showed potential for immunotherapeutic targeting of specific tumor mutations. 80% of patients had prolonged survival.[78] However, a follow-up phase III trial of rindopepimut showed no significant survival benefit.[79] Current early-stage studies targeting EGFR or its mutations are also underway, including a phase I study targeting EGFR-amplified GBM, which intends to treat patients with CART-EGFR cell therapy, and a phase I study employing combinatorial T-cell and standard-of-care therapy for EGFR-mutant GBM patients (Clinical Trial ID NCT05168423, NCT03344250). Additionally, a study of three participants treated with CARv3-TEAM-E T-cells (engineered CAR-T cells targeting EGFRvIII as well as a T-cell-engaging antibody targeting wild-type EGFR) showed rapid (but transient) tumor regression in two of three patients.[80]

There may be a biological basis for why EGFR mutation targeting may not be effective. Firstly, EGFR mutations are heterogeneous and subclonal.[81] Additionally, experiments suggest that EGFR overexpression causes preferential proliferation of mouse astrocytes vs. NSCs, suggesting that EGFR primarily impacts astrocyte-like tumor cell states rather than NSC-like states.[66] This evidence motivates research on mutational targets that also affect other cell states and potentially, other cells of origin.

These therapies may all be targeted to the SVZ region, where NSCs can be identified as the cell of origin. They offer insight into how identifying the cell of origin and mutation combination may play an important role in the future of GBM personalized therapy.

- Preventive Therapy:

Research has shown that glial progenitors (NSCs, astrocytes, OPCs) migrate from the SVZ in a predictable path that follows the developmental pathway.[34] Additionally, Korber *et al*. predicted a distant origin of GBM - up to 7 years before diagnosis – with an accumulation of specific milestone mutations, such as chromosome 7 gain, 9p loss, or 10 loss, and eventual TERT promoter mutation.[82] Targeting GBM based on precancerous indicative mutations offers a preventive treatment method for further brain infiltration and tumorigenesis. GBM's plasticity could be prevented with early detection by tracing mutations specific to the possible cells of origin.[83] Developing novel technology and methods to access precancerous mutational data through non-invasive procedures are important frontiers in research. There have already been investigations on epigenetic MRIs, which would allow imaging of the brain and its epigenetic landscape without *in vivo* sampling.[84] Identifying the patient-specific cell of origin and mutation combination may also aid in the early detection of developing tumors and allow for preventative therapies for patients at high risk of developing GBM, such as neurofibromatosis type 1 (NF1) patients.[85]

- The Tumor Microenvironment:

Mutations in growth factor genes have also been shown to be significant predictors of tumor aggressiveness, especially in the case of OPCs as the cell of origin.[86] While research has shown that a monoclonal antibody that inhibits Vascular Endothelial Growth Factor (VEGF) decreases angiogenesis, these improvements came with significant side effects.[87] Additional research and therapy must be developed to target growth factors that promote OPC self-renewal, such as PDGF inhibitors. Furthermore, depending on the cell of origin and mutation combination, the cell may interact with specific cells in the TME. For example, the PTEN mutation in GBM has been shown to recruit more macrophages.[88] Understanding these interactions could have implications for personalized therapies as well.

Conclusion

In summary, multiple lines of evidence debate the cell of origin for Glioblastoma, an aggressive brain cancer. While astrocytes have been especially controversial due to their complete differentiation and common cell markers, Neural Stem Cells and Oligodendrocyte Progenitor Cells have various convincing lines of reasoning. Liu *et al*. suggest that the cell of origin is related to the mutations accumulated and show that initial p53/NF1 mutations may occur in NSCs. Still, tumors contain OPC markers and transcriptome, thus separating the cell of mutation (the NSC in this case) from the cell of origin (OPC). Additional intricacies occur when following the mutation-dependent theory, as the cell of origin may differ by mutation, as shown by Kim *et al*., showing that GBM arises from the accumulation of driver mutations in cells of origin (NSCs and progenitor cells) and IDH mut might arise from different cells of origin that may or may not have been discovered. We conclude that the cell of mutation may differ from the cell of origin. Furthermore, we must consider differences in mutations and tumor microenvironment (growth factor, anatomical location, proximity to blood vessels).

In the future, we must determine non-invasive methods to determine each patient's cell of origin and molecular profiles to offer personalized treatment for this highly heterogeneous tumor. Additionally, logistical issues arise when catching GBM in its early stage, as it is incredibly difficult to find the singular mutated cell of origin in a sea of millions of other cells. Furthermore, it's important to consider what is realistically possible in terms of delivering treatment due to the restrictive blood-brain barrier. There are also potential limitations of targeting cells of origin; namely the possibility of off-target effects, unintended consequences of modifying progenitor cells, and the impact of GBM heterogeneity on cell-of-origin-targeted therapeutic strategies. In the future, cells of origin could be studied by directly comparing OPCs and NSCs, differentiating human induced pluripotent stem cells (hiPSCs) into OPCs and NSCs, allowing us to control the mutations introduced in an experimental setting.

Determining the cell(s) of origin of GBM is important in its potential applications to GBM treatment to improve the length and quality of life for GBM patients. In addition, identifying the cell of origin may also lead to intriguing novel strategies in the future, such as, potentially, preventing tumor cell pathogenesis, for instance, if pre-malignant cells with only a few early mutations could be targeted or prevented from acquiring further mutations

Acknowledgments

This work was supported by the Duesenberg Research Fund (to A.H.K). We thank the Kim lab members for their helpful discussion. We acknowledge Matthew Holt for creating the figures.

References

1. Grochans, S., Cybulska, A.M., Siminska, D., Korbecki, J., Kojder, K., Chlubek, D., and Baranowska-Bosiacka, I. (2022). Epidemiology of Glioblastoma Multiforme-Literature Review. Cancers (Basel) *14*. 10.3390/cancers14102412.
2. Kim, H.J., Park, J.W., and Lee, J.H. (2020). Genetic Architectures and Cell-of-Origin in Glioblastoma. Front Oncol *10*, 615400. 10.3389/fonc.2020.615400.
3. Ostrom, Q.T., Patil, N., Cioffi, G., Waite, K., Kruchko, C., and Barnholtz-Sloan, J.S. (2020). CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2013-2017. Neuro Oncol *22*, iv1-iv96. 10.1093/neuonc/noaa200.
4. Cohen-Gadol, A. (2024). Long-Term Survivors of Glioblastoma. https://www.aaroncohen-gadol.com/patients/glioma/survival/long-term-survivors.
5. Thakkar, J.P., P.; Vikram, P (2024). Glioblastoma Multiforme. https://www.aans.org/patients/conditions-treatments/glioblastoma-multiforme/.
6. Sidaway, P. (2017). CNS cancer: Glioblastoma subtypes revisited. Nat Rev Clin Oncol *14*, 587. 10.1038/nrclinonc.2017.122.
7. Lovo, E.E., Moreira, A., Barahona, K.C., Ramirez, J., Campos, F., Tobar, C., Caceros, V., Sallabanda, M., and Sallabanda, K. (2021). Stereotactic Radiosurgery for Recurrent Glioblastoma Multiforme: A Retrospective Multi-Institutional Experience. Cureus *13*, e18480. 10.7759/cureus.18480.
8. Stupp, R., Mason, W.P., van den Bent, M.J., Weller, M., Fisher, B., Taphoorn, M.J., Belanger, K., Brandes, A.A., Marosi, C., Bogdahn, U., *et al*. (2005). Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. N Engl J Med *352*, 987-996. 10.1056/NEJMoa043330.
9. Zong, H., Parada, L.F., and Baker, S.J. (2015). Cell of origin for malignant gliomas and its implication in therapeutic development. Cold Spring Harb Perspect Biol *7*. 10.1101/cshperspect.a020610.
10. Alcantara Llaguno, S.R., Wang, Z., Sun, D., Chen, J., Xu, J., Kim, E., Hatanpaa, K.J., Raisanen, J.M., Burns, D.K., Johnson, J.E., and Parada, L.F. (2015). Adult Lineage-Restricted CNS Progenitors Specify Distinct Glioblastoma Subtypes. Cancer Cell *28*, 429-440. 10.1016/j.ccell.2015.09.007.
11. Liu, C., Sage, J.C., Miller, M.R., Verhaak, R.G., Hippenmeyer, S., Vogel, H., Foreman, O., Bronson, R.T., Nishiyama, A., Luo, L., and Zong, H. (2011). Mosaic analysis with double markers reveals tumor cell of origin in glioma. Cell *146*, 209-221. 10.1016/j.cell.2011.06.014.
12. Bachoo, R.M., Maher, E.A., Ligon, K.L., Sharpless, N.E., Chan, S.S., You, M.J., Tang, Y., DeFrances, J., Stover, E., Weissleder, R., *et al*. (2002). Epidermal growth factor receptor and Ink4a/Arf: convergent mechanisms governing terminal differentiation and transformation along the neural stem cell to astrocyte axis. Cancer Cell *1*, 269-277. 10.1016/s1535-6108(02)00046-6.
13. Zhao, X., and Moore, D.L. (2018). Neural stem cells: developmental mechanisms and disease modeling. Cell Tissue Res *371*, 1-6. 10.1007/s00441-017-2738-1.
14. Purves, D.A., G; Fitzpatrick, D. (2001). Neuroscience, 2nd Edition.
15. Nottebohm, F. (1989). From bird song to neurogenesis. Sci Am *260*, 74-79. 10.1038/scientificamerican0289-74.
16. Lee, J.H., Lee, J.E., Kahng, J.Y., Kim, S.H., Park, J.S., Yoon, S.J., Um, J.Y., Kim, W.K., Lee, J.K., Park, J., *et al*. (2018). Human glioblastoma arises from subventricular zone cells with low-level driver mutations. Nature *560*, 243-247. 10.1038/s41586-018-0389-3.
17. Qin, E.Y., Cooper, D.D., Abbott, K.L., Lennon, J., Nagaraja, S., Mackay, A., Jones, C., Vogel, H., Jackson, P.K., and Monje, M. (2017). Neural Precursor-Derived Pleiotrophin Mediates Subventricular Zone Invasion by Glioma. Cell *170*, 845-859 e819. 10.1016/j.cell.2017.07.016.

DOI: 10.36838/v7i8.3

18. Matarredona, E.R., and Pastor, A.M. (2019). Neural Stem Cells of the Subventricular Zone as the Origin of Human Glioblastoma Stem Cells. Therapeutic Implications. Front Oncol *9*, 779. 10.3389/fonc.2019.00779.
19. Verdugo, E., Puerto, I., and Medina, M.A. (2022). An update on the molecular biology of glioblastoma, with clinical implications and progress in its treatment. Cancer Commun (Lond) *42*, 1083-1111. 10.1002/cac2.12361.
20. Barami, K., Sloan, A.E., Rojiani, A., Schell, M.J., Staller, A., and Brem, S. (2009). Relationship of gliomas to the ventricular walls. J Clin Neurosci *16*, 195-201. 10.1016/j.jocn.2008.03.006.
21. Loras, A., Gonzalez-Bonet, L.G., Gutierrez-Arroyo, J.L., Martinez-Cadenas, C., and Marques-Torrejon, M.A. (2023). Neural Stem Cells as Potential Glioblastoma Cells of Origin. Life (Basel) *13*. 10.3390/life13040905.
22. Chen, J., McKay, R.M., and Parada, L.F. (2012). Malignant glioma: lessons from genomics, mouse models, and stem cells. Cell *149*, 36-47. 10.1016/j.cell.2012.03.009.
23. Alcantara Llaguno, S., Chen, J., Kwon, C.H., Jackson, E.L., Li, Y., Burns, D.K., Alvarez-Buylla, A., and Parada, L.F. (2009). Malignant astrocytomas originate from neural stem/progenitor cells in a somatic tumor suppressor mouse model. Cancer Cell *15*, 45-56. 10.1016/j.ccr.2008.12.006.
24. Duan, S., Yuan, G., Liu, X., Ren, R., Li, J., Zhang, W., Wu, J., Xu, X., Fu, L., Li, Y., *et al.* (2015). PTEN deficiency reprogrammes human neural stem cells towards a glioblastoma stem cell-like phenotype. Nat Commun *6*, 10068. 10.1038/ncomms10068.
25. Yoon, S.J., Park, J., Jang, D.S., Kim, H.J., Lee, J.H., Jo, E., Choi, R.J., Shim, J.K., Moon, J.H., Kim, E.H., *et al.* (2020). Glioblastoma Cellular Origin and the Firework Pattern of Cancer Genesis from the Subventricular Zone. J Korean Neurosurg Soc *63*, 26-33. 10.3340/jkns.2019.0129.
26. Jacques, T.S., Swales, A., Brzozowski, M.J., Henriquez, N.V., Linehan, J.M., Mirzadeh, Z., C, O.M., Naumann, H., Alvarez-Buylla, A., and Brandner, S. (2010). Combinations of genetic mutations in the adult neural stem cell compartment determine brain tumour phenotypes. EMBO J *29*, 222-235. 10.1038/emboj.2009.327.
27. Tang, X., Zuo, C., Fang, P., Liu, G., Qiu, Y., Huang, Y., and Tang, R. (2021). Targeting Glioblastoma Stem Cells: A Review on Biomarkers, Signal Pathways and Targeted Therapy. Front Oncol *11*, 701291. 10.3389/fonc.2021.701291.
28. Vora, P., Venugopal, C., Salim, S.K., Tatari, N., Bakhshinyan, D., Singh, M., Seyfrid, M., Upreti, D., Rentas, S., Wong, N., *et al.* (2020). The Rational Development of CD133-Targeting Immunotherapies for Glioblastoma. Cell Stem Cell *26*, 832-844 e836. 10.1016/j.stem.2020.04.008.
29. von Bartheld, C.S., Bahney, J., and Herculano-Houzel, S. (2016). The search for true numbers of neurons and glial cells in the human brain: A review of 150 years of cell counting. J Comp Neurol *524*, 3865-3895. 10.1002/cne.24040.
30. Jakel, S., and Dimou, L. (2017). Glial Cells and Their Function in the Adult Brain: A Journey through the History of Their Ablation. Front Cell Neurosci *11*, 24. 10.3389/fncel.2017.00024.
31. Allen, N.J., and Lyons, D.A. (2018). Glia as architects of central nervous system formation and function. Science *362*, 181-185. 10.1126/science.aat0473.
32. Akay, L.A., Effenberger, A.H., and Tsai, L.H. (2021). Cell of all trades: oligodendrocyte precursor cells in synaptic, vascular, and immune function. Genes Dev *35*, 180-198. 10.1101/gad.344218.120.
33. Dawson, M.R., Polito, A., Levine, J.M., and Reynolds, R. (2003). NG2-expressing glial progenitor cells: an abundant and widespread population of cycling cells in the adult rat CNS. Mol Cell Neurosci *24*, 476-488. 10.1016/s1044-7431(03)00210-0.
34. Canoll, P., and Goldman, J.E. (2008). The interface between glial progenitors and gliomas. Acta Neuropathol *116*, 465-477. 10.1007/s00401-008-0432-9.
35. Higa, N., Akahane, T., Yokoyama, S., Yonezawa, H., Uchida, H., Takajo, T., Otsuji, R., Hamada, T., Matsuo, K., Kirishima, M., *et al.* (2022). Prognostic impact of PDGFRA gain/amplification and MGMT promoter methylation status in patients with IDH wild-type glioblastoma. Neurooncol Adv *4*, vdac097. 10.1093/noajnl/vdac097.
36. Liu, K.W., Hu, B., and Cheng, S.Y. (2011). Platelet-derived growth factor receptor alpha in glioma: a bad seed. Chin J Cancer *30*, 590-602. 10.5732/cjc.011.10236.
37. Lindberg, N., Kastemar, M., Olofsson, T., Smits, A., and Uhrbom, L. (2009). Oligodendrocyte progenitor cells can act as cell of origin for experimental glioma. Oncogene *28*, 2266-2275. 10.1038/onc.2009.76.
38. Xu, H., Zong, H., Ma, C., Ming, X., Shang, M., Li, K., He, X., Du, H., and Cao, L. (2017). Epidermal growth factor receptor in glioblastoma. Oncol Lett *14*, 512-516. 10.3892/ol.2017.6221.
39. Zong, H., Verhaak, R.G., and Canoll, P. (2012). The cellular origin for malignant glioma and prospects for clinical advancements. Expert Rev Mol Diagn *12*, 383-394. 10.1586/erm.12.30.
40. Assanah, M., Lochhead, R., Ogden, A., Bruce, J., Goldman, J., and Canoll, P. (2006). Glial progenitors in adult white matter are driven to form malignant gliomas by platelet-derived growth factor-expressing retroviruses. J Neurosci *26*, 6781-6790. 10.1523/JNEUROSCI.0514-06.2006.
41. Clavreul, S., Dumas, L., and Loulier, K. (2022). Astrocyte development in the cerebral cortex: Complexity of their origin, genesis, and maturation. Front Neurosci *16*, 916055. 10.3389/fnins.2022.916055.
42. Ghashghaei, H.T., Weimer, J.M., Schmid, R.S., Yokota, Y., McCarthy, K.D., Popko, B., and Anton, E.S. (2007). Reinduction of ErbB2 in astrocytes promotes radial glial progenitor identity in adult cerebral cortex. Genes Dev *21*, 3258-3271. 10.1101/gad.1580407.
43. Sharif, A., Legendre, P., Prevot, V., Allet, C., Romao, L., Studler, J.M., Chneiweiss, H., and Junier, M.P. (2007). Transforming growth factor alpha promotes sequential conversion of mature astrocytes into neural progenitors and stem cells. Oncogene *26*, 2695-2706. 10.1038/sj.onc.1210071.
44. Guha, A., Dashner, K., Black, P.M., Wagner, J.A., and Stiles, C.D. (1995). Expression of PDGF and PDGF receptors in human astrocytoma operation specimens supports the existence of an autocrine loop. Int J Cancer *60*, 168-173. 10.1002/ijc.2910600206.
45. Chow, L.M., Endersby, R., Zhu, X., Rankin, S., Qu, C., Zhang, J., Broniscer, A., Ellison, D.W., and Baker, S.J. (2011). Cooperativity within and among Pten, p53, and Rb pathways induces high-grade astrocytoma in adult brain. Cancer Cell *19*, 305-316. 10.1016/j.ccr.2011.01.039.
46. Sun, T., Warrington, N.M., Luo, J., Brooks, M.D., Dahiya, S., Snyder, S.C., Sengupta, R., and Rubin, J.B. (2014). Sexually dimorphic RB inactivation underlies mesenchymal glioblastoma prevalence in males. J Clin Invest *124*, 4123-4133. 10.1172/JCI71048.
47. Song, Y., Zhang, Q., Kutlu, B., Difilippantonio, S., Bash, R., Gilbert, D., Yin, C., O'Sullivan, T.N., Yang, C., Kozlov, S., *et al.* (2013). Evolutionary etiology of high-grade astrocytomas. Proc Natl Acad Sci U S A *110*, 17933-17938. 10.1073/pnas.1317026110.
48. Friedmann-Morvinski, D., Bushong, E.A., Ke, E., Soda, Y., Marumoto, T., Singer, O., Ellisman, M.H., and Verma, I.M. (2012). Dedifferentiation of neurons and astrocytes by oncogenes can induce gliomas in mice. Science *338*, 1080-1084. 10.1126/science.1226929.

49. Walz, W., and Lang, M.K. (1998). Immunocytochemical evidence for a distinct GFAP-negative subpopulation of astrocytes in the adult rat hippocampus. Neurosci Lett *257*, 127-130. 10.1016/s0304-3940(98)00813-1.
50. Zhang, Z., Ma, Z., Zou, W., Guo, H., Liu, M., Ma, Y., and Zhang, L. (2019). The Appropriate Marker for Astrocytes: Comparing the Distribution and Expression of Three Astrocytic Markers in Different Mouse Cerebral Regions. Biomed Res Int *2019*, 9605265. 10.1155/2019/9605265.
51. Ah-Pine, F., Khettab, M., Bedoui, Y., Slama, Y., Daniel, M., Doray, B., and Gasque, P. (2023). On the origin and development of glioblastoma: multifaceted role of perivascular mesenchymal stromal cells. Acta Neuropathol Commun *11*, 104. 10.1186/s40478-023-01605-x.
52. Kondo, T., and Raff, M. (2000). Oligodendrocyte precursor cells reprogrammed to become multipotential CNS stem cells. Science *289*, 1754-1757. 10.1126/science.289.5485.1754.
53. Kondo, T., and Raff, M. (2004). Chromatin remodeling and histone modification in the conversion of oligodendrocyte precursors to neural stem cells. Genes Dev *18*, 2963-2972. 10.1101/gad.309404.
54. Tissot, T., Ujvari, B., Solary, E., Lassus, P., Roche, B., and Thomas, F. (2016). Do cell-autonomous and non-cell-autonomous effects drive the structure of tumor ecosystems? Biochim Biophys Acta *1865*, 147-154. 10.1016/j.bbcan.2016.01.005.
55. Zong, H., Espinosa, J.S., Su, H.H., Muzumdar, M.D., and Luo, L. (2005). Mosaic analysis with double markers in mice. Cell *121*, 479-492. 10.1016/j.cell.2005.02.012.
56. Verhaak, R.G., Hoadley, K.A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M.D., Miller, C.R., Ding, L., Golub, T., Mesirov, J.P., *et al*. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell *17*, 98-110. 10.1016/j.ccr.2009.12.020.
57. Cancer Genome Atlas Research, N. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature *455*, 1061-1068. 10.1038/nature07385.
58. Wang, Z., Sun, D., Chen, Y.J., Xie, X., Shi, Y., Tabar, V., Brennan, C.W., Bale, T.A., Jayewickreme, C.D., Laks, D.R., *et al*. (2020). Cell Lineage-Based Stratification for Glioblastoma. Cancer Cell *38*, 366-379 e368. 10.1016/j.ccell.2020.06.003.
59. Wang, Q., Hu, B., Hu, X., Kim, H., Squatrito, M., Scarpace, L., deCarvalho, A.C., Lyu, S., Li, P., Li, Y., *et al*. (2017). Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment. Cancer Cell *32*, 42-56 e46. 10.1016/j.ccell.2017.06.003.
60. Zhang, P., Xia, Q., Liu, L., Li, S., and Dong, L. (2020). Current Opinion on Molecular Characterization for GBM Classification in Guiding Clinical Diagnosis, Prognosis, and Therapy. Front Mol Biosci *7*, 562798. 10.3389/fmolb.2020.562798.
61. Wang, J., Cazzato, E., Ladewig, E., Frattini, V., Rosenbloom, D.I., Zairis, S., Abate, F., Liu, Z., Elliott, O., Shin, Y.J., *et al*. (2016). Clonal evolution of glioblastoma under therapy. Nat Genet *48*, 768-776. 10.1038/ng.3590.
62. Fedele, M., Cerchia, L., Pegoraro, S., Sgarra, R., and Manfioletti, G. (2019). Proneural-Mesenchymal Transition: Phenotypic Plasticity to Acquire Multitherapy Resistance in Glioblastoma. Int J Mol Sci *20*. 10.3390/ijms20112746.
63. Kim, Y., Varn, F.S., Park, S.H., Yoon, B.W., Park, H.R., Lee, C., Verhaak, R.G.W., and Paek, S.H. (2021). Perspective of mesenchymal transformation in glioblastoma. Acta Neuropathol Commun *9*, 50. 10.1186/s40478-021-01151-4.
64. Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L., *et al*. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science *344*, 1396-1401. 10.1126/science.1254257.
65. Suva, M.L., and Tirosh, I. (2020). The Glioma Stem Cell Model in the Era of Single-Cell Genomics. Cancer Cell *37*, 630-636. 10.1016/j.ccell.2020.04.001.
66. Neftel, C., Laffy, J., Filbin, M.G., Hara, T., Shore, M.E., Rahme, G.J., Richman, A.R., Silverbush, D., Shaw, M.L., Hebert, C.M., *et al*. (2019). An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. Cell *178*, 835-849 e821. 10.1016/j.cell.2019.06.024.
67. Richards, L.M., Whitley, O.K.N., MacLeod, G., Cavalli, F.M.G., Coutinho, F.J., Jaramillo, J.E., Svergun, N., Riverin, M., Croucher, D.C., Kushida, M., *et al*. (2021). Gradient of Developmental and Injury Response transcriptional states defines functional vulnerabilities underpinning glioblastoma heterogeneity. Nat Cancer *2*, 157-173. 10.1038/s43018-020-00154-9.
68. Bergmann, N., Delbridge, C., Gempt, J., Feuchtinger, A., Walch, A., Schirmer, L., Bunk, W., Aschenbrenner, T., Liesche-Starnecker, F., and Schlegel, J. (2020). The Intratumoral Heterogeneity Reflects the Intertumoral Subtypes of Glioblastoma Multiforme: A Regional Immunohistochemistry Analysis. Front Oncol *10*, 494. 10.3389/fonc.2020.00494.
69. Fernandes, C., Costa, A., Osorio, L., Lago, R.C., Linhares, P., Carvalho, B., and Caeiro, C. (2017). Current Standards of Care in Glioblastoma Therapy. In Glioblastoma, S. De Vleeschouwer, ed. 10.15586/codon.glioblastoma.2017.ch11.
70. Platten, M. (2017). EGFRvIII vaccine in glioblastoma-In-ACT-IVe or not ReACTive enough? Neuro Oncol *19*, 1425-1426. 10.1093/neuonc/nox167.
71. Stupp, R., Taillibert, S., Kanner, A., Read, W., Steinberg, D., Lhermitte, B., Toms, S., Idbaih, A., Ahluwalia, M.S., Fink, K., *et al*. (2017). Effect of Tumor-Treating Fields Plus Maintenance Temozolomide vs Maintenance Temozolomide Alone on Survival in Patients With Glioblastoma: A Randomized Clinical Trial. JAMA *318*, 2306-2316. 10.1001/jama.2017.18718.
72. Nathanson, D.A., Gini, B., Mottahedeh, J., Visnyei, K., Koga, T., Gomez, G., Eskin, A., Hwang, K., Wang, J., Masui, K., *et al*. (2014). Targeted therapy resistance mediated by dynamic regulation of extrachromosomal mutant EGFR DNA. Science *343*, 72-76. 10.1126/science.1241328.
73. Ramirez, Y.P., Weatherbee, J.L., Wheelhouse, R.T., and Ross, A.H. (2013). Glioblastoma multiforme therapy and mechanisms of resistance. Pharmaceuticals (Basel) *6*, 1475-1506. 10.3390/ph6121475.
74. Persson, A.I., Petritsch, C., Swartling, F.J., Itsara, M., Sim, F.J., Auvergne, R., Goldenberg, D.D., Vandenberg, S.R., Nguyen, K.N., Yakovenko, S., *et al*. (2010). Non-stem cell origin for oligodendroglioma. Cancer Cell *18*, 669-682. 10.1016/j.ccr.2010.10.033.
75. Jiang, Y., Marinescu, V.D., Xie, Y., Jarvius, M., Maturi, N.P., Haglund, C., Olofsson, S., Lindberg, N., Olofsson, T., Leijonmarck, C., *et al*. (2017). Glioblastoma Cell Malignancy and Drug Sensitivity Are Affected by the Cell of Origin. Cell Rep *19*, 1080-1081. 10.1016/j.celrep.2017.04.053.
76. Gravina, G.L., Mancini, A., Colapietro, A., Vitale, F., Vetuschi, A., Pompili, S., Rossi, G., Marampon, F., Richardson, P.J., Patient, L., *et al*. (2017). The novel CXCR4 antagonist, PRX177561, reduces tumor cell proliferation and accelerates cancer stem cell differentiation in glioblastoma preclinical models. Tumour Biol *39*, 1010428317695528. 10.1177/1010428317695528.
77. Olympios, N., Gilard, V., Marguet, F., Clatot, F., Di Fiore, F., and Fontanilles, M. (2021). TERT Promoter Alterations in Glioblastoma: A Systematic Review. Cancers (Basel) *13*. 10.3390/cancers13051147.

DOI: 10.36838/v7i8.3

78. Reardon, D.A., Desjardins, A., Vredenburgh, J.J., O'Rourke, D.M., Tran, D.D., Fink, K.L., Nabors, L.B., Li, G., Bota, D.A., Lukas, R.V., *et al.* (2020). Rindopepimut with Bevacizumab for Patients with Relapsed EGFRvIII-Expressing Glioblastoma (ReACT): Results of a Double-Blind Randomized Phase II Trial. Clin Cancer Res *26*, 1586-1594. 10.1158/1078-0432.CCR-18-1140.
79. Weller, M., Butowski, N., Tran, D.D., Recht, L.D., Lim, M., Hirte, H., Ashby, L., Mechtler, L., Goldlust, S.A., Iwamoto, F., *et al.* (2017). Rindopepimut with temozolomide for patients with newly diagnosed, EGFRvIII-expressing glioblastoma (ACT IV): a randomised, double-blind, international phase 3 trial. Lancet Oncol *18*, 1373-1385. 10.1016/S1470-2045(17)30517-X.
80. Choi, B.D., Gerstner, E.R., Frigault, M.J., Leick, M.B., Mount, C.W., Balaj, L., Nikiforow, S., Carter, B.S., Curry, W.T., Gallagher, K., and Maus, M.V. (2024). Intraventricular CARv3-TEAM-E T Cells in Recurrent Glioblastoma. N Engl J Med *390*, 1290-1298. 10.1056/NEJMoa2314390.
81. Furnari, F.B., Cloughesy, T.F., Cavenee, W.K., and Mischel, P.S. (2015). Heterogeneity of epidermal growth factor receptor signalling networks in glioblastoma. Nat Rev Cancer *15*, 302-310. 10.1038/nrc3918.
82. Korber, V., Yang, J., Barah, P., Wu, Y., Stichel, D., Gu, Z., Fletcher, M.N.C., Jones, D., Hentschel, B., Lamszus, K., *et al.* (2019). Evolutionary Trajectories of IDH(WT) Glioblastomas Reveal a Common Path of Early Tumorigenesis Instigated Years ahead of Initial Diagnosis. Cancer Cell *35*, 692-704 e612. 10.1016/j.ccell.2019.02.007.
83. De Silva, M.I., Stringer, B.W., and Bardy, C. (2023). Neuronal and tumourigenic boundaries of glioblastoma plasticity. Trends Cancer *9*, 223-236. 10.1016/j.trecan.2022.10.010.
84. Lam, F., Chu, J., Choi, J.S., Cao, C., Hitchens, T.K., Silverman, S.K., Liang, Z.P., Dilger, R.N., Robinson, G.E., and Li, K.C. (2022). Epigenetic MRI: Noninvasive imaging of DNA methylation in the brain. Proc Natl Acad Sci U S A *119*, e2119891119. 10.1073/pnas.2119891119.
85. Alcantara Llaguno, S.R., and Parada, L.F. (2016). Cell of origin of glioma: biological and clinical implications. Br J Cancer *115*, 1445-1450. 10.1038/bjc.2016.354.
86. Bogler, O., Wren, D., Barnett, S.C., Land, H., and Noble, M. (1990). Cooperation between two growth factors promotes extended self-renewal and inhibits differentiation of oligodendrocyte-type-2 astrocyte (O-2A) progenitor cells. Proc Natl Acad Sci U S A *87*, 6368-6372. 10.1073/pnas.87.16.6368.
87. Kong, D.H., Kim, M.R., Jang, J.H., Na, H.J., and Lee, S. (2017). A Review of Anti-Angiogenic Targets for Monoclonal Antibody Cancer Therapy. Int J Mol Sci *18*. 10.3390/ijms18081786.
88. Chen, P., Zhao, D., Li, J., Liang, X., Li, J., Chang, A., Henry, V.K., Lan, Z., Spring, D.J., Rao, G., *et al.* (2019). Symbiotic Macrophage-Glioma Cell Interactions Reveal Synthetic Lethality in PTEN-Null Glioma. Cancer Cell *35*, 868-884 e866. 10.1016/j.ccell.2019.05.003

Authors

Grace Qian is a current Stanford undergraduate student interested in neuroscience, immunology, and oncology. She is studying human biology on the pre-med track.

IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Awareness and Educational Gaps in Color Vision Deficiency: A Survey-Based Analysis from Gujarat, India

Freya Prajapati[1], Aahan Prajapati[1], Ajay Goyal, PhD[2]

1) Adani International School, Adani Shantigram, Ahmedabad, Gujarat 382421; freyaprajapati2510@gmail.com
2) Institute of Design, Nirma University, Ahmedabad, Gujarat 382481 India

ABSTRACT: Color blindness (or Color Vision Deficiency (CVD)) affects millions of people worldwide, impacting the ability to differentiate colors. This study investigated how many people in Gujarat, India, are aware of the consequences of CVD. A survey conducted on 482 individuals (including students, teachers, parents, and professionals) revealed unawareness and a lack of school testing policies. Participants supported awareness programs, better teaching materials, and smartphone apps for CVD individuals. Logistic regression and decision trees are promising techniques to identify key factors influencing awareness. The results highlight the need for more testing, education, and support in Gujarati schools.

KEYWORDS: Color Vision Deficiency, Awareness Survey, logistic regression, Decision Tree Classification.

Introduction

Color blindness or Color Vision Deficiency (CVD) is a prevalent genetic illness.[1] Most individuals remain unaware that a family member may be affected by CVD, leading to delayed identification and support.[2] Awareness of CVD at the initial stage is limited. This affects everyday life and professional environments.[3] In India, the issue is underreported and underexplored. Limited literature is available that investigates awareness of color blindness. No study was found related to children. Ahmad *et al*. found that commercial drivers often ignore their color blindness.[4] Harikrishnan and Mohan similarly observed a lack of CVD awareness among professionals in the visual media field in Kerala, India.[5] Dhaliwal *et al*. noted that while some doctors perceive CVD as a risk, others support inclusive policies through proper training.[6]

Technology offers new opportunities to support colorblind individuals. AI-powered tools and mobile applications can assist with color identification.[7] Salih *et al*. introduced wearable devices that enhance color perception in time.[8] Tigwell worked on inclusive design, addressing CVD awareness gaps among product developers.[9] Despite technological advances, schools in India still lack structured screening policies for CVD, especially in developing countries like India. Senjam *et al*. found limited access to assistive technologies, even in specialized schools.[10] Pinner highlighted the need for adaptations for CVD students.[11] Cohen *et al*. further revealed that people often do not notice missing colors in their vision, underscoring how CVD can go undetected.[12]

This study surveys awareness levels, knowledge gaps, and available spaces for colorblind individuals in Indian municipal schools of Gujarat. It builds on the epidemiological work by Fareed *et al*. and newer studies emphasizing the role of education, policy, and assistive technology.[1] This study is based on the hypothesis that awareness of CVD among school communities is predominantly associated with individual roles, age, prior exposure to color confusion, and familiarity with diagnostic tools such as the Ishihara test. Objectives of the study are as follows:

- To assess awareness of CVDs among students, teachers, parents, and professionals.
- To understand the testing and accommodating policies of schools.
- Examine the psychological and social impact.
- Analyze career challenges.
- Identify key demographic and behavioral factors influencing awareness.

Methods

A Google survey was done in 319 Gujarat municipal schools, collecting responses from 482 individuals (172 females, Figure 1(a)): 274 students, 63 teachers, 53 parents, and 92 professionals responded to the form (Figure 1(b)). A convenience sampling strategy was used to select participants. Participation was voluntary, and survey links were distributed via school administrators and online educational groups. Participants varied in age group, classified as 0-10, 10-20, 20-30, 30-40, 40-50, 50-60, 60-70, and 70-80 years. Major participants were in the age group of 10-50 years old (Figure 1(c)). The questionnaire consisted of 30 items, including multiple-choice questions. It was divided into four sections: demographic details, general awareness of CVD, personal or familial experience with CVD, and perspectives on educational support and career impact. Statistical analyses were applied to ascertain significant relationships between age, role, color confusion, and other variables with the levels of awareness. A logistic regression model was developed to predict support for school counseling and awareness programs based on role, awareness of CVD, and experiences with face color confusion. A decision tree classifier was also trained using Gini impurity, a max depth of 5, and a minimum split of 10 samples to determine influencing factors for awareness. All statistical analyses and machine learning models were implemented using MATLAB R2024b.

DOI: 10.36838/v7i8.4

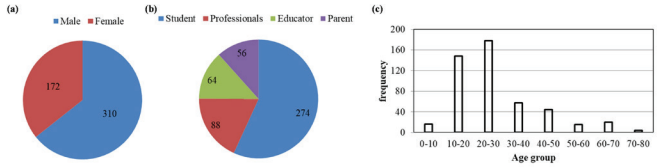


Figure 1: (a) Gender distribution showing counts of male and female participants, (b) participant roles categorized as student, professional, educator, and parent, and (c) age group distribution of all respondents represented across predefined age intervals. The participant pool was predominantly male students. Most respondents were under 30 years of age.

Result and Discussion

The poll results show modest knowledge levels; 66.4% of respondents had heard of CVDs, but just 33.4% knew their frequency. 37.1% of respondents knew that most cases include red-green color deficit; awareness of the Ishihara test was 42.6%. 91.3% have never been tested for CVD (Figure 2). 12.6% of participants have a CVD family member (Figure 3). 66.6% of participants agreed that CVD reduces self-confidence, particularly among students (Figure 4). 60.2% of affected patients said they had never witnessed taunting due to CVD (Figure 4). Respondents expressed strong worry about career challenges caused by CVD. While 38.5% supported alternate work assessments to accommodate a CVD individual, 30.2% agreed that job rejection due to CVDs was appropriate. 73.6% of participants favored changes in teaching materials and examination techniques. 65.5% supported teacher training programs, while 87.2% supported school counseling activities for CVD students. 86.9% of respondents had never utilized color blindness assistance, while 75.5% were interested in a color identification smartphone application.

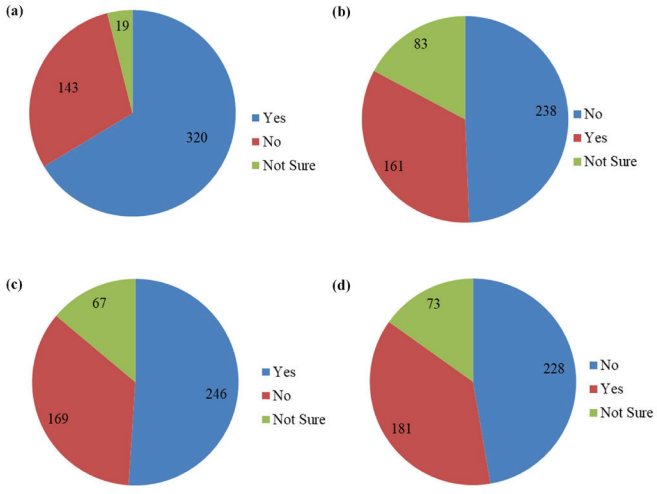


Figure 2: Participant responses on various aspects of Color Vision Deficiency (CVD): (a) general awareness of color blindness, (b) knowledge of its prevalence, (c) familiarity with the Ishihara Test and school screening procedures, and (d) understanding of the broader impact of CVD. Most participants were aware of color blindness, but significant gaps existed in understanding its prevalence, testing methods, and broader impact, highlighting the need for enhanced education and awareness efforts.

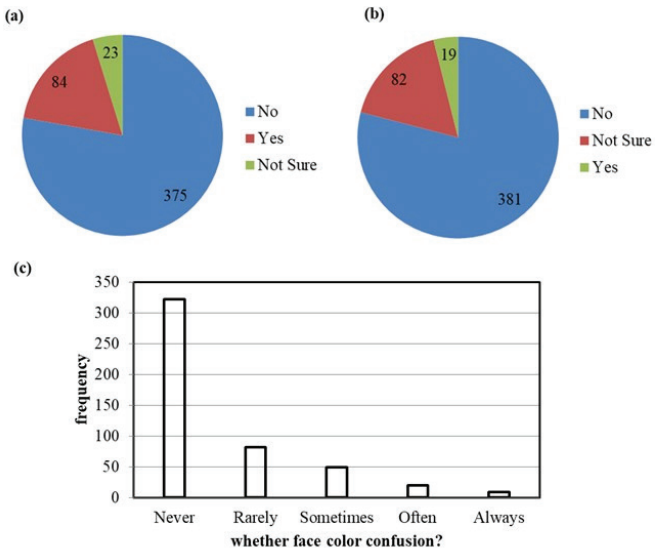


Figure 3: Statistics on participant experiences related to Color Vision Deficiency (CVD): (a) whether the participant has undergone CVD testing, (b) the presence of CVD in any family member, and (c) instances of color confusion experienced by the participant in daily activities. Most participants reported not being tested for CVD and were unaware of any family history of the condition. Additionally, the majority stated they never experience color confusion in daily life, which suggests limited direct engagement with CVD-related challenges.

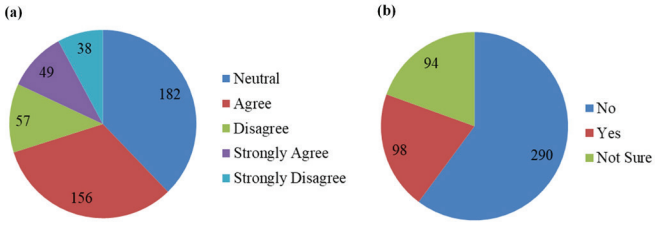


Figure 4: (a) Participant opinions on the impact of Color Vision Deficiency (CVD) on self-confidence, and (b) responses regarding observations of teasing or bullying related to CVD at school or in the workplace. Participants agreed that CVD can impact self-confidence, although many remained neutral. Most respondents reported not witnessing teasing or bullying related to CVD.

Statistical analyses showed that younger respondents (<30 years) had significantly higher awareness (t-test, p=0.018), awareness varied by profession (ANOVA, p=0.0026), age had a weak negative correlation with awareness (-0.15), role had a moderate association with awareness (Cramer's V=0.29), and gender and job disqualification opinions had weak associations with awareness (0.18 and 0.16, respectively).

The Logistic Regression model, with 76.8% accuracy, has been given in Eq. (1). The decision tree classifier achieved an accuracy of 72.16%, with the best performance in predicting color blindness awareness (Yes class, 86% recall). The most influential factors were awareness of Ishihara tests, prior color blindness testing, and family history. The most influential factors are given in Figure 5.

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = -1.78 + 0.42(Role) + 1.12(Awareness) + 0.68(Color\ Confusion) \quad (1)$$

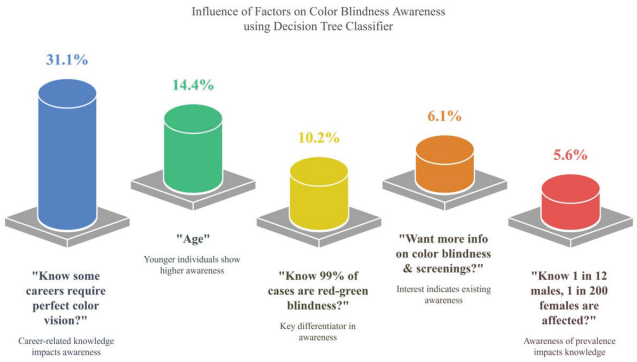


Figure 5: Influencing factors for Color Vision Deficiency (CVD) awareness identified using a Decision Tree Classifier. The figure illustrates the factors that influence selected variables in predicting awareness of color blindness, including career-related knowledge, age, understanding of red-green color blindness, interest in screenings, and awareness of prevalence statistics. Career-related knowledge is the most influential factor in predicting CVD awareness, followed by age and understanding red-green blindness.

This study has a few limitations. The data was collected from municipal schools in Gujarat, which may not represent the broader population across India or globally. All responses were self-reported, possibly introducing biases or inaccuracies in understanding and experiences. The study did not include clinical testing to confirm CVD, relying on participants' awareness and perception. Additionally, the use of online surveys may have excluded individuals without internet access or digital literacy, potentially limiting the diversity of the sample.

■ Conclusion

This study inspected the awareness of CVD patients from Gujarat's municipal schools. Among all 482 participants, 66.4% heard of CVD, and 33.4% understood its prevalence. Most participants have not had them tested or even heard of the associated tests (for example, the Ishihara test). Students, teachers, and parents have limited knowledge about the impact of CVD on daily life. Most participants with CVD have acknowledged that the condition affects their confidence and limits their career choices. This study suggested a need for teacher training, school counseling, color-friendly materials, and mobile apps. Machine learning models (Logistic regression and decision tree models) exhibited that awareness and personal experience were key determinants of support for CVD awareness programs. Though performed in Gujarat schools, this study highlighted an urgent requirement for early screening, awareness campaigns, and revised education policies.

■ References

1. Fareed, M.; Anwar, M. A.; Afzal, M. Prevalence and Gene Frequency of Color Vision Impairments among Children of Six Populations from North Indian Region. *Genes & Diseases* **2015**, *2* (2), 211–218. https://doi.org/10.1016/j.gendis.2015.02.006.
2. Robertson, D. L. Invisibility in the Colorblind Era: Examining Legitimized Racism against Indigenous Peoples. *American Indian Quarterly* **2015**, *39* (2), 113–153. https://doi.org/10.5250/amerindiquar.39.2.0113.
3. Ullucci, K.; Battey, D. Exposing Color Blindness/Grounding Color Consciousness: Challenges for Teacher Education. *Urban Education* **2011**, *46* (6), 1195–1225. https://doi.org/10.1177/0042085911413150.
4. Ahmad, A.; Yadav, P.; Pant, K.; Tripathi, A.; Dubey, G. Reviewing the Prevalence of Refractive Errors and Color Blindness among Commercial Drivers. *International Journal Of Community Medicine And Public Health* **2024**, *11* (8), 3309–3313. https://doi.org/10.18203/2394-6040.ijcmph20242195.
5. Mohan, G.; Harikrishnan, D. INVESTIGATING THE PREVALENCE, AWARENESS AND COPING MECHANISMS FOR COLORBLINDNESS BY THE MALAYALAM VISUAL MEDIA PROFESSIONALS. *JOURNAL OF THE ASIATIC SOCIETY OF MUMBAI* **2024**, *27* (3), 1–10.
6. Dhaliwal, U.; Singh, S.; Nagpal, G.; Kakkar, A. Perceptions of Specialist Doctors of the Ability of Doctors with Colour Vision Deficiency to Practise Their Specialty Safely. *Indian Journal of Medical Ethics* **2020**, *5* (3), 268–277.
7. Pendhari, N.; Shaikh, D.; Shaikh, N.; Nagori, A. Comprehensive Color Vision Enhancement for Color Vision Deficiency: A Tensorflow and Keras Based Approach. *ICTACT Journal on Image and Video Processing* **2024**, *14* (4), 3282–3292. https://doi.org/10.21917/ijivp.2024.0467.
8. Salih, A. E.; Elsherif, M.; Ali, M.; Vahdati, N.; Yetisen, A. K.; Butt, H. Ophthalmic Wearable Devices for Color Blindness Management. *Advanced Materials Technologies* **2020**, *5* (8), 1901134. https://doi.org/10.1002/admt.201901134.
9. Tigwell, G. W. Nuanced Perspectives Toward Disability Simulations from Digital Designers, Blind, Low Vision, and Color Blind People. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*; CHI '21; Association for Computing Machinery: New York, NY, USA, 2021; pp 1–15. https://doi.org/10.1145/3411764.3445620.
10. Senjam, S. S.; Manna, S.; Vashist, P.; Gupta, V.; Grover, S.; Kumar, V. A.; Titiyal, J. S. Improving Assistive Technology Access to Students with Low Vision and Blindness in Delhi: A School-Based Model. *Indian J Ophthalmol* **2023**, *71* (1), 257–262. https://doi.org/10.4103/ijo.IJO_1281_22.
11. Pinner, M. The Effect of Training upon Faculty Stages of Concern about Making Color Vision Deficiency Adaptations. Ph.D. thesis, Liberty University, Lynchburg, VA 24515, 2021. https://digitalcommons.liberty.edu/doctoral/2877.
12. Cohen, M. A.; Botch, T. L.; Robertson, C. E. The Limits of Color Awareness during Active, Real-World Vision. *Proceedings of the National Academy of Sciences* **2020**, *117* (24), 13821–13827. https://doi.org/10.1073/pnas.1922294117.

■ Authors

Freya Prajapati is a Grade 9 student with a deep interest in biology and veterinary sciences. As the Co-founder of Aiding Colors, she actively participates in the colorblindness screening camps. She has surveyed 482 individuals to understand the awareness and accommodation in schools for colorblind students. Seeing her brother, Aahan, struggle to differentiate between colors motivated her to use technology to help people with invisible disabilities.

Aahan Prajapati is a Grade 11 student with a deep passion for computational biology, focusing on computational vision science and visual perception modeling. Being red-green colorblind, he aims to develop innovative healthcare solutions that can positively impact people's lives with his project, Aiding Colors – An initiative for screening and counseling for colorblindness.

DOI: 10.36838/v7i8.4

Dr. Ajay Goyal is a PhD graduate in Computational Biomechanics of Mice's Bones. He has done extensive work in bone adaptation, occupant modeling, fracture healing, implant design, injury analysis, skin cancer detection, and color blindness diagnosis. He serves at Nirma University as a research faculty member and exam coordinator.

IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

A Comparative Analysis of AI Model and Traditional Model in 4D Trajectory Prediction Using Different Error Metrics

Harshitha Vishnu

Fremont High School, 575 W Fremont Ave, Sunnyvale, California, 94087, USA; mailtoharshithavishnu@gmail.com
Mentor: Bahae Samhan

ABSTRACT: The development of advanced predictive models can help ensure safe and efficient operations within the complex area of the air traffic domain. Accurate 4-dimensional (latitude, longitude, altitude, and time) trajectory prediction plays a crucial role in enhancing the safety and efficiency of modern air traffic management systems. This study presents a comparative analysis of two trajectory prediction models, a linear regression model and a neural network, focusing on their ability to accurately predict the position of an aircraft given factors such as latitude, longitude, altitude, and time. Using data obtained from OpenSky, both models were evaluated for prediction accuracy on different error metrics like Euclidean error, MAE, MSE, RMSE, MAPE, and R^2. Results from the study indicate that while the neural network model performed better on the Euclidean error, the linear model had lower scores on all the other error metrics. This underperformance may be attributed to inadequate feature engineering, overfitting, or insufficient hyperparameter tuning, and implementing techniques like hyperparameter tuning methods, regularization methods, or additional features could enhance the model's accuracy. These findings provide critical insights into the strengths and limitations of each model and highlight the importance of balancing model complexity with performance requirements to refine predictive systems in next-generation ATM systems.

KEYWORDS: Robotics and Intelligent Machines, Machine Learning, 4D trajectory prediction, Aircraft Trajectory Modeling, Artificial Intelligence in Aviation.

Introduction

The need for accurate trajectory prediction models in civil aviation is essential for improving flight safety, enhancing operational efficiency, and minimizing delays.[1] As global air traffic continues to increase, precise models are required to optimize flight paths, allowing for the effective use of increasingly congested airspace.[2] One such advancement is the use of 4D trajectory prediction, which not only accounts for latitude, longitude, and altitude but also incorporates time, making it superior to traditional 3D predictions, which generally include latitude, longitude, and altitude.[3] In this context, latitude refers to the aircraft's north-south position while longitude refers to its east-west position relative to the Earth's equator, altitude measures the distance above a surface or sea level, and time tracks positional changes over a period.

The implementation of accurate trajectory prediction models can aid in making efficient use of limited airspace. This way, flights can follow the optimal flight paths while increasing the number of aircraft in an area, while the area is still safe. These optimal paths also reduce the time required for aircraft to reach their destinations while also ensuring the safety of the passengers and the flight crew.[2]

Traditional trajectory prediction methods, such as Linear Regression, have been widely used for identifying linear relationships in flight data in small datasets.[4] However, this model is not capable of predicting non-linear relationships often present in aviation data, so it might underfit the data. This would lead to a lot of outliers, resulting in a higher residual error, which means the model would make inaccurate predictions.[5]

Artificial Intelligence and Machine Learning are promising alternatives to aid in trajectory prediction. AI models, especially neural networks, excel at identifying non-linear relationships in data.[6] Neural networks have multiple layers with multiple nodes, and they produce an output through multiplication, summation, activation function application, and other such processes. These models can identify patterns in complex data that go unnoticed by simpler Regression models.[7]

This study addresses the following research question: How does a neural network-based AI model compare to traditional Linear Regression in predicting 4D aircraft trajectories in terms of accuracy and practical applicability for air traffic management? This research aims to compare the predictive accuracy of a traditional Linear Regression model with that of a Neural Network model, evaluating how AI can improve 4D trajectory modeling in civil aviation. Key metrics for comparison include Euclidean Error, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and R-squared (R^2) scores. The data used for this analysis was obtained from OpenSky, a free platform offering real-time aircraft position data (including latitude, longitude, altitude, and time) collected through ADS-B transponders. The dataset includes variables such as call sign, origin country, time position, last contact, longitude, and latitude. For this study, the input variables were longitude, latitude, altitude, and time deltas. To compare the differences in models' predictive accuracy, a Linear regression model was employed as the traditional model, and a Neural Network model with TensorFlow and TFlite was integrated as the AI model.

DOI: 10.36838/v7i8.5

This study contributes to civil aviation by emphasizing the benefits of comparing traditional and AI-based models for trajectory prediction and aims to provide insights into which methodologies best optimize flight paths and manage air traffic, improving the overall efficiency of air traffic management systems. By addressing the limitations of traditional models and demonstrating how enhanced prediction models can streamline aviation systems, the research supports more efficient use of airspace, reduces delays, and ensures safer skies for the future.

■ Background and Evolution of 4D Trajectory Modeling

4D trajectory modeling, which refers to the management of aircraft positions in three dimensions, namely latitude, longitude, and altitude, in addition to time, has evolved significantly over the years. This development is primarily driven by the rapid growth of the aviation industry and reduced flight intervals, which prioritize the need to optimize routes for safety, efficiency, and environmental impact.

The initial efforts in trajectory prediction for civil aviation were grounded in basic physical models, such as kinematic equations of motion. These early models primarily used deterministic methods to predict future aircraft positions based on historical data.[8] However, they were simplistic and did not account for various factors like weather, wind, or air traffic.

Research in 4D trajectory prediction began with relatively simpler models. One of the first techniques that emerged for trajectory prediction was linear regression, where trajectory data for aircraft was fitted to a linear model. This allowed for simple predictions based on past observations or an easily recognizable pattern. However, this method's accuracy was limited due to its inability to identify complex patterns in nonlinear data, which meant it often underfitted the data and did not provide good predictions.[8]

Recently, with the development of the aviation industry, statistical models have become prominent. Among the most significant advancements was the introduction of the Kalman Filter in trajectory modeling.[9] The Kalman filter is an algorithm that can use data that is noisy or inaccurate and estimate an unknown variable with greater accuracy. Kalman filters surpassed linear regression in air trajectory prediction due to their ability to handle dynamic systems and uncertainties effectively.[10]

As computing power grew in the 1990s, more sophisticated nonlinear models were introduced.[11] Instead of merely predicting future positions, these models integrated the concept of 4D trajectory modeling, incorporating not only three spatial dimensions but also time. Methods such as the Extended Kalman Filters (EKF), Monte Carlo Simulations, and stochastic models gained popularity during this time in the air traffic management industry.[12]

Today, 4D trajectory modeling has become advanced. With the advancement of deterministic models, which provide a single, concise prediction of an aircraft's future trajectory based on initial data and well-defined equations of motion. Model Predictive Control (MPC), which predicts the future state of an aircraft based on data about its current state, is a great advancement in deterministic models because it continuously optimizes to surpass limitations like weather and air traffic. Advancements in deterministic kinematic models include more sophisticated aircraft dynamics and flight mechanics.

Research is increasingly focusing on models that can adjust predictions dynamically as real-time data, like weather updates or unexpected air traffic changes, becomes available, improving model responsiveness.[13,14] Additionally, combining deterministic and probabilistic approaches, such as deterministic models enhanced with neural networks, is a growing area, as it can offer more reliable predictions by balancing precision with flexibility for handling unpredictable factors.[15]

Probabilistic models are statistical models that use uncertainty as a factor in their predictions. By considering all the possibilities, these models can deal with otherwise unpredictable changes like weather, air traffic congestion, and unexpected delays. Some examples of probabilistic models include Monte Carlo simulations and Bayesian networks.[12] In Monte Carlo simulations, multiple trajectories are predicted when various factors like weather and wind speed are inputted into the model. Bayesian networks represent the relationships between different variables probabilistically, which is useful in incorporating both historical and real-time data.

Despite their statistical flexibility, traditional deterministic and probabilistic models still struggle with the dynamic, nonlinear complexities of real-world aviation. Deterministic approaches rely on fixed equations, limiting their adaptability to unforeseen disruptions like sudden weather shifts or air traffic congestion.[13] Probabilistic methods, while accounting for uncertainty, often require extensive computational resources to scale effectively and may lack the granularity to capture intricate spatial-temporal patterns.[12] These shortcomings highlight the need for more adaptive, data-driven solutions, prompting the exploration of AI-based techniques, which leverage machine learning to dynamically refine predictions using real-time data.

Traditional trajectory prediction models are useful in many situations, but also include a lot of limitations. These models are unable to handle relationships between non-linear data, like wind speed and turbulence.[16] If they are asked to find patterns in complex, non-linear data, they would simplify the output and therefore underfit the data. These models also have limited adaptability to constantly updating real-time data. These foundational models also perform poorly in complex or noisy data. They are not adapted to handle the complexity required for large-scale data.

Collectively, traditional trajectory prediction models—whether deterministic or probabilistic—face inherent constraints. Key limitations include reliance on simplified linear assumptions, leading to underfitting in nonlinear scenarios; poor scalability with high-dimensional, noisy data; and static architectures unable to assimilate real-time updates efficiently. These gaps necessitate a paradigm shift toward AI-driven models capable of learning complex patterns autonomously and adjusting predictions dynamically.

Neural networks, particularly deep learning architectures, have shown promising results in predicting 4D trajectories. One of the most common architectures for 4D trajectory prediction is the Long-Term short-memory network. LTSM networks are a type of recurrent neural network (RNN) that excels in mastering long-term dependencies.[17] LTSMs contain a memory cell that can hold information for a long period of time. LTSMs can also decide what information to include and what to remove. LTSMs are effective at learning from historical trajectory data and predicting future positions in both space and time, and this makes them ideal predictors of the aircraft's future trajectory. Graph Neural Networks (GNNs) are an emerging architecture for aviation trajectory prediction, especially in multi-aircraft scenarios. GNNs contain data points called nodes, which are linked by lines called edges. GNNs model the interactions between different aircraft as nodes in a graph, with edges representing potential interactions (such as proximity with other aircraft or potential conflicts).

This progression from traditional deterministic models to AI-driven 4D trajectory modeling reflects the industry's shift toward more adaptive and predictive solutions that are capable of addressing the complexities of modern aviation. By incorporating real-time data and handling nonlinear interactions, AI models offer unprecedented accuracy and flexibility that traditional models cannot achieve, setting a new standard in trajectory prediction.[8,18]

The advancements in trajectory modeling, from deterministic methods to AI-driven techniques, demonstrate the aviation industry's commitment to addressing the complexities of modern air traffic management. While traditional models like linear regression have laid the foundation for trajectory prediction, their limitations in handling nonlinear and dynamic data necessitate the exploration of more sophisticated approaches. This motivates this research to investigate how integrating a neural network-based AI model can improve the accuracy and efficiency of 4D trajectory predictions compared to traditional methods. The following sections outline the methodology, including data collection and modeling approaches, that underpin this comparative analysis.

Methodology

Data Collection and Description:

The dataset used for this research was obtained from OpenSky, a free receiver network that contains credible aircraft information received through ADS-B transponders. The attributes extracted from the dataset include Callsign, Origin country, time position, last contact, latitude, longitude, and altitude. As 4D trajectory prediction involves latitude, longitude, altitude, and time, the corresponding attributes will be significant to the research. In aviation, a callsign is a unique identifier that consists of the aircraft's name along with a combination of unique numbers or letters. This is used to distinguish an aircraft during communication with the control center or other aircraft, which avoids confusion when multiple aircraft are in the same space. These attributes contribute to accurate tracking of the aircraft's position, enabling precise trajectory prediction. For this research, the callsign will be used to distinguish the aircraft, which is critical in predicting its trajectory. Additionally, attributes like origin, country, and last contact also help identify a particular aircraft and distinguish it from others.

Model Selection:

To address the need for accurate trajectory prediction, two modeling approaches were introduced: a traditional linear regression model and a neural network-based AI model. Each approach offers unique strengths and leverages different aspects of the data. The linear regression model provides a straightforward approach to understanding simpler relationships among variables, while the AI model can handle the non-linear complexities in location data, which may enhance flexibility and prediction accuracy for real-time predictions.

Linear regression attempts to predict the relationship between an independent variable (Time delta, Latitude delta, and Longitude delta) and the target variables (Latitude and Longitude), assuming a linear relationship between these variables. This model also assumes constant variance, independence of errors, no multicollinearity, and performance.

The AI model's hyperparameters (e.g., layer size, activation functions, epochs) were selected based on empirical best practices for trajectory prediction tasks, balancing computational efficiency and predictive performance. While systematic hyperparameter optimization techniques like grid search or cross-validation were not employed due to resource constraints, key design choices were validated through iterative testing.

The AI model consists of an input layer, two hidden layers with 64 nodes and ReLU activation functions, and an output layer with two nodes to predict latitude and longitude. ReLU (Rectified Linear Unit) is used in the hidden layers because it is able to handle nonlinear data effectively, and the output layer remains linear to directly predict latitude and longitude. The model utilizes 64 nodes in each layer to achieve a balanced learning capacity without overfitting the data. The output layer contains two nodes corresponding to the latitude and longitude predictions. TensorFlow was chosen for training because of its ability to execute various tasks across many platforms and its general performance with neural networks after training. The model was converted to TensorFlow Lite (TFLite) for its compactness, efficiency, and ability to deploy the model on mobile or embedded systems, which allows for faster prediction time, ideal for real-time trajectory prediction.

For model training and evaluation, the training dataset is split into 80% training and 20% testing for both models using TrainTestSplit. 10% of the training data is further split to be used as validation data during neural network training to monitor and reduce overfitting. For the neural network, a default batch size was assigned following the framework's optimal settings for efficiency. However, adjustments can be made depending on hardware and computational power. The neural network is trained for 10 epochs, balancing training time with achieving sufficient accuracy in learning trajectory patterns.

DOI: 10.36838/v7i8.5

Evaluation metrics:

Euclidean Error measures the direct spatial distance between predicted and true positions, providing a clear indication of geographic accuracy, which is essential for trajectory applications. A smaller Euclidean error indicates that two points are closer together, while a larger error means they are further apart, and in this scenario, a smaller error would be optimal. This metric is very easy to interpret, and its calculations are straightforward. Euclidean Error is particularly suitable for trajectory prediction because it directly quantifies positional deviations in physical space—a critical requirement for aviation safety and navigation. Additionally, its simplicity ensures computational efficiency, making it practical for real-time systems where rapid error assessment is necessary. Euclidean Error is defined as:

$$Euclidean\ Error = \sqrt{\left(Latitude_{pred} - Latitude_{true}\right)^2 + \left(Longitude_{pred} - Longitude_{true}\right)^2}$$

Equation (1)

Mean Absolute Error (MAE) calculates the average absolute error in predictions, giving insight into general prediction accuracy and robustness against outliers. It gives an idea of how much, on average, the predictions deviate from the actual values. MAE gives the average magnitude of errors in the same units as the output. The lower the MAE, the better the model is at making accurate predictions. MAE can be represented as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|$$

Equation (2)

Where yi is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of data points.

Mean Squared Error (MSE) is the average of the squared differences between predicted and actual values. It penalizes larger errors more than smaller ones due to the squaring of the differences. The lower the MSE, the better the model is performing. A small MSE indicates the model is making mostly accurate predictions. It is more sensitive to outliers because it emphasizes larger errors. The formula for MSE is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

Equation (3)

The square root of MSE, Root Mean Squared Error (RMSE), has the same units as the target, making it easier to interpret in terms of geographical units. It provides insight into the overall prediction accuracy, emphasizing larger errors. Like MSE, the lower the RMSE, the better the model's performance. However, RMSE is easier to interpret since it is in the same units as the data. RMSE is especially useful when larger errors need to be penalized more. RMSE can be derived as shown:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}$$

Equation (4)

Mean Absolute Percentage Error (MAPE) measures error as a percentage, making it helpful for understanding the model's relative accuracy and comparing different models' performance. A lower MAPE means better predictive performance. It is particularly useful when comparing the accuracy of models across different datasets or scales. However, it can give extreme values when actual values are close to zero. It is also easily interpretable as a percentage. The formula that can be used to derive MAPE is:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_1}\right|$$

Equation (5)

R-squared (R^2) indicates the proportion of variance in the data explained by the model. Its values range from 0 to 1. Higher R^2 values show that the model explains more of the observed data variance, which is especially useful for assessing linear models' effectiveness. Negative values can occur if the model performs worse than a simple baseline model. R-squared can be represented as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

Equation (6)

Where $\bar{y}$ is the mean of the actual values

Experimental Setup:

This project utilizes Python for implementation, with libraries including TensorFlow, TFLite, and Scikit-Learn. TensorFlow is used to build, train, and convert the model into a TFLite model for efficiency, and Scikit-Learn, which contains modules like StandardScaler and TrainTestSplit, is used for linear regression and data scaling. The TensorFlow model was initially trained in a typical Python runtime environment compatible with TensorFlow 2.x. The TFLite runtime interpreter is lightweight, making it suitable for resource-constrained environments, ensuring efficient model predictions even for systems with limited computational power, such as the home PC used to implement the model for this research. An 80/20 train-test split is applied, with 80% of the data used for model training and 20% for final evaluation to assess the model's overall performance. Within the training data, a 10% validation split is applied to monitor the neural network's performance during training. This approach helps track model convergence and identify any potential overfitting to the training data. Given the model's relatively simple architecture and the deterministic nature of trajectory data (e.g., temporal dependencies), targeted hyperparameter selection was prioritized over exhaustive search methods. This approach aligned with the goal of lightweight deployment for real-time applications. Finally, feature scaling was applied, using StandardScaler, to improve model generalization and performance.

■ Results

This section compares the performance of the traditional linear model with that of the AI model, using a range of key error metrics to assess prediction accuracy. The primary metric applied, Euclidean Error, measures the straight-line distance between predicted and actual values, providing an overall indicator of model accuracy. Additionally, several other metrics were utilized to offer a nuanced performance assessment. Mean Absolute Error (MAE) was calculated to provide the average magnitude of errors in the same units as the target variable, giving a straightforward measure of typical error size. Mean Squared Error (MSE), which represents the average

squared differences between predicted and actual values, was also applied. This metric penalizes larger errors more heavily than smaller ones due to the squaring of the differences, making it sensitive to outliers and emphasizing significant deviations in prediction. Root Mean Squared Error (RMSE), similar to MSE, also accentuates larger errors but is easier to interpret because it is expressed in the same units as the target data. To provide insights into performance in relative terms, Mean Absolute Percentage Error (MAPE) was used, which expresses error as a percentage, making it particularly useful for understanding model accuracy across different scales. Lastly, the R^2 score, or coefficient of determination, was applied to evaluate how well each model explains the variance in the target data. This metric indicates the proportion of variance in the dependent variable that is predictable from the independent variables, thus providing an estimate of each model's explanatory power. Collectively, these metrics offer a comprehensive view of model performance, revealing strengths and limitations across various aspects of predictive accuracy and error sensitivity.

While the AI model achieved a lower Euclidean error, indicating better performance on this metric, it underperformed on other key metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and R^2.

The traditional linear model and the AI model displayed notable differences in predicting latitude. Although the AI model achieved a Euclidean Error of 36.073, which was substantially lower than the traditional model's Euclidean Error of 59.538, the AI model's performance was inconsistent across other metrics. For instance, the MAE for the AI model was 1835.694, which is considerably higher than the traditional model's MAE of 10.986, indicating a greater overall prediction error. Similar trends were observed in the MSE, RMSE, and MAPE values, with the AI model recording markedly higher errors (MSE of 3369773.141 and RMSE of 1835.694) compared to the traditional model (MSE of 223.239 and RMSE of 14.941). Additionally, while the traditional model had an R2 score of 0.458, the AI model's error could not be computed, represented by NaN, meaning "Not a Number," which signifies an undefined or unrepresented value. This shows how a valid error could not be provided for this metric. These discrepancies suggest that while the AI model achieves closer proximity to actual values in specific instances (as indicated by lower Euclidean Error), it may not consistently capture the underlying data structure as effectively as the traditional model across all error metrics. Table 1 and Figure 1 summarize these results for latitude prediction.

Table 1: Comparison in latitude prediction accuracy between the Traditional and the AI model using various error metrics. The results indicate that the traditional model performed better than the AI model on every metric except for Euclidean error, where the AI model performed better by a small margin.

	Traditional Model	AI Model
Euclidean Error	59.538	36.073
MAE	10.986	1835.694
MAP	223.239	3369773.141
RMSE	14.941	1835.694
MAPE	53.768	5861.486
R^2	0.458	NaN

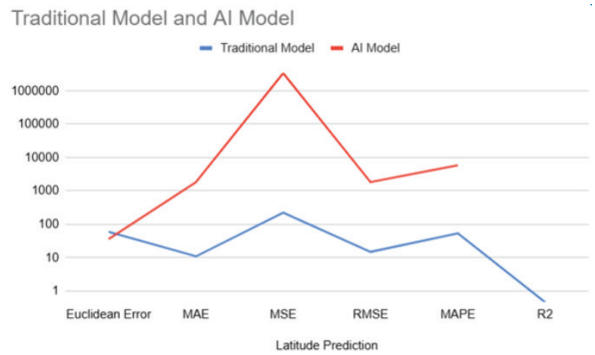


Figure 1: The visual representation of the models' performance in predicting the latitude indicates that the AI model performed worse than the traditional model at every error metric except Euclidean error (a lower error equals better accuracy). It can also be seen that there is a large margin of difference between the traditional model's and AI model's performance for each metric.

For longitude prediction, the AI model again demonstrated a lower Euclidean Error of 36.073 compared to the traditional model's 59.538. However, as with latitude prediction, the AI model underperformed significantly on other metrics. The MAE for the AI model was 5259.041, substantially higher than the traditional model's 57.724, and the MSE and RMSE values for the AI model were also significantly higher (MSE of 27657509.561 and RMSE of 5259.041) compared to the traditional model's MSE of 4560.720 and RMSE of 67.533. Furthermore, the MAPE value for the AI model was exceedingly high at 5428.508, contrasting sharply with the traditional model's MAPE of 102.830. The R^2 metric could not be computed for the AI model (NaN), whereas the traditional model achieved an R^2 of 0.472, suggesting some level of explanatory power in the traditional model. Table 2 and Figure 2 display these comparative results for longitude prediction.

Table 2: Comparison in longitude prediction accuracy between the Traditional and the AI model using various error metrics. The results indicate that the traditional model performed better than the AI model on every metric except for Euclidean error, where the AI model performed better by a small margin.

	Traditional Model	AI Model
Euclidean Error	59.538	36.073
MAE	57.724	5259.041
MAP	4560.720	27657509.561
RMSE	67.533	5259.041
MAPE	102.830	5428.508
R^2	0.472	NaN

DOI: 10.36838/v7i8.5

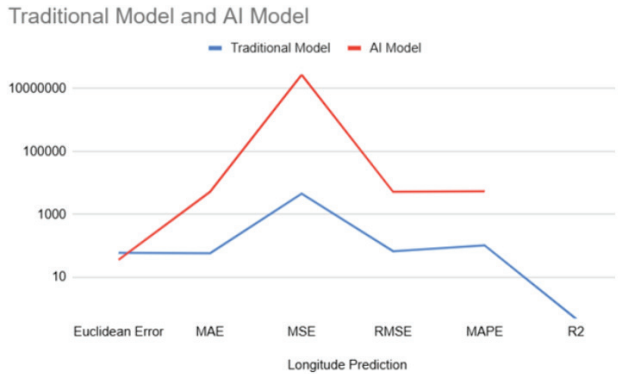


Figure 2: The visual representation of the models' performance in predicting the longitude indicates that the AI model performed worse than the traditional model at every error metric except Euclidean error (a lower error equals better accuracy). It can also be seen that there is a large margin of difference between the traditional model's and AI model's performance for each metric, suggesting the traditional model may be reliable for generalized local predictions.

The findings indicate that while the AI model excels in minimizing Euclidean Error, it struggles with other critical error metrics, particularly MAE, MSE, RMSE, and MAPE, for both latitude and longitude predictions. This discrepancy may stem from the AI model's sensitivity to data inconsistencies or its potential overfitting to specific patterns within the dataset. Conversely, the traditional linear model demonstrated more consistent performance across a wider range of error metrics, suggesting it may be more reliable for generalized location predictions despite a slightly higher Euclidean Error.

These results highlight the importance of selecting appropriate performance metrics when evaluating predictive models, as different metrics can highlight varying aspects of model accuracy and reliability. Future studies may explore ways to optimize the AI model's performance across all error metrics, potentially by refining model architecture or employing additional data preprocessing techniques to improve generalization.

Discussion

The performance analysis shows that although the AI model performed better than the Linear model at Euclidean Error, it performed worse in other errors. This may be because while Euclidean error can be effective at assessing overall spatial accuracy in predictions, it is not able to capture the nuances of individual coordinate predictions, such as latitude or longitude.[19] For instance, integrating spatial autocorrelation in machine learning models can enhance accuracy by accounting for geographical data. However, relying only on Euclidean distance can result in individual errors being masked by overall distance calculations, leading to potentially misleading performance assessments. Traditional models like spatial lag, which incorporate spatial features, can better capture the spatial dependency and reduce prediction errors.[18,20]

The AI model's poorer performance on MAE, MSE, and RMSE (despite its Euclidean error advantage) suggests it struggles with coordinate-specific precision. Unlike Euclidean distance, which aggregates errors into a single spatial value, these metrics penalize directional biases (e.g., consistent overestimation of altitude). The AI model's focus on holistic spatial accuracy may come at the cost of localized errors in individual dimensions (latitude/longitude/altitude), which are weighted equally in traditional metrics.[21]

Another reason the traditional model performed better than the AI model could be the fact that traditional models often perform better on simpler metrics like RMSE, MSE, and MAE.[22] This provides stable and interpretable results. If the model had a simpler structure, such as in linear regression or traditional random forest models, it delivered more consistent results due to fewer complexities in data relationships. While AI models may introduce noise or fail to generalize due to their complexity, traditional models, such as spatial lag or linear regression, prioritize interpretability, making them more reliable for simpler datasets or metrics like RMSE and MAE.[23] This stability is particularly evident when spatial dependencies are weak and the models are not overfitted. As seen in the comparison between spatial models and traditional machine learning models, simpler models tend to generalize better under specific conditions.[18] Furthermore, traditional models generally require fewer computational resources, enabling faster optimization and fewer errors caused by resource limitations during training.[24]

Additionally, AI models may exhibit significant errors due to several factors, such as overfitting training data, inadequate feature engineering, or insufficient hyperparameter tuning. Overfitting is particularly problematic when the model learns noise in the training data, leading to poor generalization on test data.[25] Especially on small or imbalanced datasets, AI models are prone to overfitting and memorizing instead of generalizing, while traditional models, which are less data-intensive, are less likely to overfit.[26] Additionally, the inclusion of irrelevant or insufficient features may increase the runtime and hinder the model's ability to capture complex relationships between inputs and outputs.[25] Furthermore, fine-tuning model architecture and parameters is crucial to optimizing performance. AI models often depend on well-designed features. Inadequate preprocessing or irrelevant features can degrade their performance compared to traditional models, which are more robust to such shortcomings.[27] For instance, AI models may overlook natural spatial relationships unless specifically programmed, unlike spatial regression models designed to address geographic dependencies explicitly.[28] The AI model's suboptimal performance on non-Euclidean metrics may also stem from its inability to prioritize coordinate-specific errors. Euclidean distance aggregates spatial deviations into a single value, potentially masking directional biases.[29] In contrast, traditional models optimize for individual coordinate errors directly, aligning better with metrics like MAE. Additionally, the AI model's fixed architecture (e.g., 64-node hidden layers) may lack the adaptability to capture localized spatial patterns, whereas traditional methods like spatial lag explicitly model geographic dependencies.[30] This limitation becomes pronounced when training data lacks sufficient variability in spatial-temporal features, further exacerbating directional errors. While the AI model's superior performance in Euclidean

error suggests potential for spatial accuracy, its inconsistency across other metrics raises questions about its robustness. For example, probabilistic models like Bayesian networks16 or hybrid approaches[16] may better handle uncertainty in dynamic conditions, such as weather disruptions or air traffic variability, by combining deterministic predictions with probabilistic adjustments.

The AI model's higher MAE/RMSE scores may also stem from its inability to prioritize error types critical for aviation. For example, altitude errors (safety-critical) and time errors (delay-sensitive) are treated equally with lateral position errors in the loss function. Traditional models, by contrast, often optimize for domain-specific priorities (e.g., penalizing altitude deviations more heavily), aligning better with operational needs.

Fundamentally, the AI model's failures in non-Euclidean metrics reflect a misalignment between its training objective (minimizing bulk spatial error) and aviation's need for dimension-aware precision. While Euclidean optimization suits general spatial tasks, trajectory prediction requires balancing heterogeneous errors (e.g., time vs. altitude), necessitating custom loss functions or hybrid architectures that blend AI's nonlinear capacity with traditional models' interpretable constraints.

It is found that advanced AI techniques like deep learning, though powerful, require careful design choices to minimize these issues.[15,18] Advancing the performance of AI models requires the exploration of more sophisticated models or alternative architectures. Emerging techniques such as reinforcement learning, generative adversarial networks (GANs), transfer learning, and neuro evolution offer promising solutions to issues that traditional machine learning models struggle with. In addition, these models significantly enhance predictive accuracy and reduce errors, leading to better performance. For instance, GANs have been utilized in scenarios requiring creative data generation, such as vehicle trajectory prediction,[13] and show great potential for use in the trajectory prediction of aircraft. Similarly, reinforcement learning allows models to interact with their environments, developing accuracy over time, which could improve the performance of the model, leading to better trajectory predictions. Explainable AI (XAI), which refers to a set of processes and methods designed to make AI models more transparent and interpretable to humans, also offers frameworks that could not only improve performance but also make AI decisions more interpretable and transparent, aiding in debugging and model refinement. It takes accountability for its decisions while also mitigating bias.[31]

Moreover, emerging technologies, such as quantum AI, are beginning to demonstrate significant potential, offering massive computational power capable of addressing highly complex tasks with greater precision.[32] This diversity in AI techniques provides a range of alternatives that may better align with the specific requirements of a given problem domain and dataset characteristics. These advancements highlight the importance of exploring and adopting innovative architectures, particularly in scenarios where traditional models underperform.[14]

The findings of this study align with the potential of advanced techniques like reinforcement learning (RL) and GANs to address observed shortcomings. RL's iterative reward-based optimization could dynamically adjust predictions in response to real-time errors (e.g., wind shifts), while GANs could synthesize rare but critical scenarios (e.g., extreme turbulence) to improve generalization. Explainable AI (XAI) frameworks, such as SHAP or LIME, could further bridge the gap between the AI model's 'black-box' predictions and the interpretability of traditional models, enabling targeted debugging of coordinate-specific errors (e.g., latitude bias) and fostering trust in aviation applications.

The linear model's stability in MAE and RMSE shows us its reliability for scenarios where computational efficiency is prioritized, indicating the need to align model selection with specific operational requirements. For example, AI models for precision in controlled environments versus traditional models for generalizability. Therefore, we suggest that future work should test hybrid frameworks to mitigate the limitations observed in standalone AI or linear approaches.

Limitations and Future Directions

While the data obtained from the OpenSky network was accurate, it was restricted in quantity, potentially limiting the model's ability to generalize across diverse scenarios. The dataset also had a few inconsistencies, such as missing values, which impacted the predictive accuracy of the model. Inaccuracies in recorded flight parameters or limited temporal and spatial resolution may have introduced noise into the training process, which might have led the model to provide biased or inaccurate predictions. The complexity of the AI model used in this study might have contributed to its overfitting or underfitting of the data. Overfitting arises when the model captures noise in the training data, resulting in reduced generalization for unseen data. In contrast, underfitting occurs when a model is too simple to capture the complexities or the underlying patterns in the data. To address these issues, hyperparameter tuning methods like grid search, random search, or Bayesian optimization could have been implemented to find the best hyperparameter configurations. Additionally, regularization methods like weight decay, dropout, or early stopping could have been utilized to reduce overfitting. On the other hand, to minimize underfitting, the complexity of the model could have been increased using feature engineering and hyperparameter tuning. Moreover, the choice of features utilized might not have fully captured the dynamics of the system being modeled. Features like weather, turbulence, and air traffic at that specific time could have been incorporated to capture underlying patterns and increase predictive accuracy. While the data collected represents a diverse range of flights from many flight regions, the data was collected during a specific time frame, so it might not be representative of data obtained during other time frames, which limits the model's ability to accurately predict data from other time frames. The dataset may also contain inherent biases based on geographical coverage, aircraft types, or airline operators represented in the OpenSky network. Such

DOI: 10.36838/v7i8.5

sampling biases could skew the model's understanding of typical flight patterns. Additionally, missing values in critical parameters like altitude or velocity may reflect systemic gaps in ADS-B coverage rather than random noise. These data quality issues compound the challenges of training reliable predictive models, as they may cause the AI system to learn artifacts of data collection limitations rather than true trajectory patterns.

Several data preprocessing and augmentation approaches could mitigate these limitations. For missing values, multiple imputation techniques could estimate plausible values while accounting for uncertainty, rather than simple deletion or mean imputation. For temporal biases, implementing time-based stratification during train-test splits would ensure all time periods are represented. For spatial biases, geographic weighting could balance representation across flight regions. To reduce noise, Kalman filtering could be utilized to smooth erratic position reports while preserving true trajectory patterns.

For more robust testing, implementing k-fold cross-validation (e.g., 5- or 10-fold) is recommended, especially for the traditional linear regression model.[33] This technique provides a more comprehensive performance estimate across different data partitions. Additionally, using early stopping or monitoring validation loss can help prevent overfitting in the neural network model by stopping training when improvement plateaus.[34]

To enhance AI model performance, exploring advanced methodologies is crucial. Techniques such as reinforcement learning, GANs, and transfer learning offer promising avenues for improvement. For example, GANs have shown success in generating synthetic trajectory data to improve predictive models,[13] while reinforcement learning can iteratively optimize models by interacting with dynamic environments.

Explainable AI (XAI) frameworks also present significant opportunities, making AI predictions more transparent and interpretable. By improving model accountability and mitigating bias, XAI frameworks not only enhance performance but also increase trust in AI systems.[31]

Emerging technologies, such as quantum AI, offer unparalleled computational power, which could revolutionize trajectory prediction by addressing the complexities of large-scale, nonlinear datasets with higher precision.[32] Exploring these techniques alongside traditional models will help identify the most effective methodologies for optimizing 4D trajectory prediction.

These findings have direct implications for real-world air traffic management (ATM) systems. The comparative performance analysis suggests that hybrid systems combining traditional models' reliability with AI's pattern recognition capabilities could optimize trajectory prediction in operational environments. For instance, linear models could serve as baseline predictors while AI components handle complex, nonlinear scenarios like weather disruptions or congested airspace, creating a more robust ensemble system. These results could inform the phased implementation of AI in ATM systems. The performance metrics established here provide concrete benchmarks for aviation authorities evaluating prediction systems, particularly in balancing accuracy requirements with computational constraints. Future work could test these models in simulation environments mirroring actual air traffic control workflows to validate operational feasibility.

In summary, this study shows that while traditional models excel in simplicity, efficiency, and interpretability, AI models show immense promise for handling complex, nonlinear data. However, their effectiveness depends on overcoming challenges such as overfitting and inadequate feature engineering. The integration of advanced techniques, including GANs, reinforcement learning, and XAI, highlights the potential for AI to set new standards in 4D trajectory modeling, improving air traffic management and operational efficiency.

Acknowledgments

The author would like to acknowledge Prof. Bahae Samhan, PhD, Associate Professor of Business Information Systems at Illinois State University, for his constant support and guidance throughout the research process.

References

1. Shi, Z., Xu, M., Pan, Q., Yan, B., & Zhang, H. (2018, July). LSTM-based flight trajectory prediction. In *2018 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE*.
2. Rosenow, J., Lindner, M., & Scheiderer, J. (2021). Advanced flight planning and the benefit of in-flight aircraft trajectory optimization. Sustainability, 13(3), 1383.
3. Tian, Y., He, X., Xu, Y., Wan, L., & Ye, B. (2020). 4D trajectory optimization of commercial flight for green civil aviation. *IEEE Access*, 8, 62815-62829.
4. Shin, S., Lee, S., Son, C., Yee, K. (2022). Aircraft design parameter estimation using data-driven machine learning models. In *Proceedings of the 33rd Congress of the International Council of the Aeronautical Sciences* (ICAS 2022).
5. Li, C., & Braun, M. (2022). Analyzing aircraft turnaround processes using small dataset relationships.
6. Calvo Fernández, E., Cordero, J. M., Vouros, G., Pelekis, N., Kravaris, T., Georgiou, H., Fuchs, G., Andrienko, N., Andrienko, G., Casado, E., Scarlatti, D., & Costas, P. (2017). DART: A machine-learning approach to trajectory prediction and demand-capacity balancing
7. Zhao, L., Zhang, J., & Li, L. (2019). Trajectory forecasting with neural networks: An empirical evaluation and a new hybrid model. *IEEE Transactions on Intelligent Transportation Systems*, 20(6), 2095-2106. https://doi.org/10.1109/TITS.2018.2879827
8. Zeng, W., Chu, X., Xu, Z., Liu, Y., & Quan, Z. (2022). Aircraft 4D Trajectory Prediction in Civil Aviation: A Review. *Aerospace,* 9(2), 91.
9. Welch, G., & Bishop, G. (2006). An introduction to the Kalman filter
10. Prevost, C. G., Desbiens, A., & Gagnon, E. (2007). Extended Kalman filter for state estimation and trajectory prediction of a moving object detected by an unmanned aerial vehicle. In *Proceedings of the IEEE International Conference on Control Applications* (pp. 1058-1063). IEEE.
11. Bondarenko, D., & Baskin, K. (2017). Big History, Complexity Theory, and Life in a Non-Linear World. From Big Bang to Galactic Civilizations: *A Big History Anthology*, 3, 183-196.
12. Murphy, K. P. (2022). *Probabilistic machine learning: an introduction*. MIT Press.
13. Xu, T., Zhang, Q., Liu, Y., & Wang, X. (2023). Adaptive real-time 4D trajectory prediction in the presence of dynamic weather and

air traffic conditions. *IEEE Transactions on Intelligent Transportation Systems*.
14. Pang, Y., Liu, X., Zhang, M., & Huang, Y. (2023). Real-time adaptability in 4D trajectory prediction under dynamic conditions. *Journal of Air Transport Management*, 102, 114798. https://doi.org/10.1016/j.jairtraman.2023.114798
15. Ma, J., & Tian, L. (2023). Predictive modeling for aircraft trajectories based on hybrid CNN-LSTM networks. *IEEE Transactions on Intelligent Transportation Systems*, 24(7), 1326–1341. https://doi.org/10.1109/TITS.2023.3034893
16. Zhang, Y., Zhang, D., & Jiang, H. (2023). Review of Challenges and Opportunities in Turbulence Modeling: A Comparative Analysis of Data-Driven Machine Learning Approaches. *Journal of Marine Science and Engineering*, 11(7), 1440. https://doi.org/10.3390/jmse11071440
17. Staudemeyer, R., & Morris, E. (2024). – Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks. Ar5iv. https://ar5iv.org/html/1909.09586
18. Ma, L., Meng, X., & Wu, Z. (2024). Data-Driven 4D Trajectory Prediction Model Using Attention-TCN-GRU. Aerospace, 11(4), 313.
19. O. E. Drummond, X. R. Li, and C. He, "Comparison of Various Static Multiple-Model Estimation Algorithms," in *Proc. 1998 SPIE Conf. on Signal and Data Processing of Small Targets*, vol. 3373, pp. 510–527, 1998.
20. Naser, M.Z., Alavi, A.H. Error Metrics and Performance Fitness Indicators for Artificial Intelligence and Machine Learning in Engineering and Sciences. Archit. Struct. Constr. 3, 499–517 (2023). https://doi.org/10.1007/s44150-021-00015-8
21. Alligier, R., Gianazza, D., & Durand, N. (2023). Machine Learning and Trajectory Prediction: Why Euclidean Metrics Mislead. Journal of Air Transportation, 31(2), 45-59.
22. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
23. Bai, Y., Yu, W., & Li, H. (2023). Evaluating spatial data models: Metrics and their implications for geographic predictions. *Journal of Spatial Science*, 68(1), 12–29.
24. Zhang, L., & Han, P. (2023). Tuning hyperparameters for deep learning: A systematic approach to optimization. Computational Intelligence, 39(3), 345–360
25. Aliferis, C., & Simon, G. (2024). Overfitting, Underfitting, and General Model Overconfidence and Under-Performance Pitfalls and Best Practices in Machine Learning and AI. In *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences: Best Practices and Pitfalls* (pp. 477-524). Cham: Springer International Publishing.
26. Wang, R., & Chen, Y. (2023). The role of feature engineering in addressing data dependency issues in machine learning. *Expert Systems with Applications*, 226, 120079
27. Liu, X., Zhang, Q., & Wang, J. (2023). Comparing machine learning and traditional statistical models: A metric-based performance analysis. *International Journal of Data Science*, 12(2), 145–162.
28. Lee, J., & Shin, K. (2023). Handling scalability and overfitting in deep learning for big data analysis. *Journal of Big Data*, 10(1), 45–61
29. Besse, P. C., Guillouet, B., Loubes, J. M., & François, R. (2016). "Review of nonlinear trajectory prediction methods for aircraft." *Journal of Aerospace Information Systems*, 13(3), 123-138
30. Nguyen, D. T., Le, L., & Sharma, R. (2021). "Limitations of deep learning for spatial data: A case study in aviation trajectory prediction." *IEEE Transactions on Intelligent Transportation Systems*, 22(5), 2765-2776
31. Letrache, K., & Ramdani, M. (2023). *Explainable Artificial Intelligence: A Review and Case Study on Model-Agnostic Methods*. IEEE. https://ieeexplore.ieee.org/abstract/document/10373722
32. Wichert, A. (2020). P*rinciples of quantum artificial intelligence: Quantum problem solving and machine learning*. World Scientific. https://doi.org/10.1142/9789811224317_0001
33. Wong, T. T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross-validation. Pattern Recognition, 48(9), 2839–2846. https://doi.org/10.1016/j.patcog.2015.03.009
34. Liu, X.; Kounadi, O.; Zurita-Milla, R. Incorporating Spatial Autocorrelation in Machine Learning Models Using Spatial Lag and Eigenvector Spatial Filtering Features. ISPRS Int. J. Geo-Inf. 2022, 11, 242. https://doi.org/10.3390/ijgi11040242

Author

Harshitha Vishnu is a freshman at Fremont High School in Sunnyvale, California. She has a deep passion for STEM, especially in the fields of Aerospace Engineering and Artificial Intelligence. She aims to pursue a future in AI and aerospace engineering, with a focus on solving real-world challenges through technology and innovation.

DOI: 10.36838/v7i8.5

IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

The Intersection of MVP Culture and MLB Revenue

Michael R. Bronshteyn

Calabasas High School, 22855 West Mulholland Hwy, Calabasas, CA, 91302, USA; Mbronshteyn070@gmail.com
Mentor: Jason Damm

ABSTRACT: This study analyzes the impact of a Major League Baseball (MLB) player's Most Valuable Player (MVP) season on their team's financial performance, specifically focusing on gate receipts and attendance metrics. Utilizing data across multiple seasons from MLB teams, the study employs statistical analysis through data gathered in a panel regression to determine the correlation between a player's MVP season and subsequent changes in these financial metrics. The findings suggest that an MVP season can significantly increase a team's gate receipts and attendance, as fans are drawn to witness exceptional performances. Further study could extend these findings by gauging the long-term financial impact of an MVP season, including how merchandise sales, media rights, and sponsorships are changed in subsequent seasons. Furthermore, investigating factors such as the economic environment, fan demographics, and team market size that impact the financial outcomes of an MVP season would offer even more detail in the development of these financial results. This study's dataset, covering the 2009-2019 timeframe, does not account for more contemporary trends or external economic factors like recessions, which can influence fan spending behavior.
KEYWORDS: Mathematics, Panel Regression, MVP Season, Gate Receipts, Attendance.

Introduction

In Major League Baseball (MLB), a player's Most Valuable Player (MVP) season can have a profound impact on a team's financial performance, specifically on gate receipts and attendance. An MVP season, marked by remarkable individual achievements, not only demonstrates elite skill but also drives significant economic growth for the team. The draw of witnessing talent firsthand often increases fan interest, leading to higher attendance at games. Fans are drawn by the opportunity to see a player performing at their peak, creating a heightened sense of excitement around each game.[1]

A panel regression model was used to analyze the relationship between MVP seasons and the dependent variables. A panel regression combines cross-sectional data (multiple teams) and time-series data (multiple seasons), allowing for the analysis of how changes in MVP status affect financial metrics over time. This method is ideal for this study because it accounts for both variations across teams and the passage of time while controlling for unobserved factors that could influence outcomes, such as team location, market size, or long-standing fan bases. By employing fixed effects, the model isolates the true impact of an MVP season on the financial vitality of MLB teams.[2]

To better understand this dynamic, the following sections will explore historical data on attendance and gate receipts before, during, and after MVP seasons across several MLB teams. Through this analysis, the goal is to quantify the financial impact of an MVP player on an MLB franchise.

Literature Review

The economic impact of star players, particularly during standout seasons, is a well-established phenomenon in sports economics, often referred to as the "Superstar Effect." Rosen first articulated this effect, positing that a small group of exceptionally talented individuals, or superstars, amass a disproportionate share of market attention and revenue.[1] His theory suggests that fans are uniquely drawn to witness extraordinary talent, creating heightened demand and a higher willingness to pay for tickets to games featuring these players, thereby leading to significant financial returns for teams with star athletes.

Studies in both the NBA and MLB demonstrate the substantial role star players play in driving attendance and revenue. Hausman and Leonard found that dominant players in the NBA, like Larry Bird, could increase game attendance by up to 50%, indicating that fans are not solely motivated by team success but are also attracted to individual performance excellence.[3] This reflects fans' desire to see players with elite skill sets in action, a factor that can sustain attendance and drive revenue even during seasons of suboptimal team performance. Similarly, Scully showed in MLB that teams with MVP-caliber players experience significant boosts in ticket sales as fans willingly pay premium prices to watch these standout players perform.[4]

The presence of a star player in MLB, particularly during an MVP season, is shown to have a lasting impact on fan turnout and revenue. This aligns with the findings of Berri *et al.*, who demonstrated that NBA teams with marquee players maintained strong attendance levels, largely due to the intrinsic draw of individual star power.[5]

Furthermore, the impact of an MVP season can have a lasting impact on a team's revenue, often strengthening a team's market positioning and revenue in subsequent years. Humphreys and Johnson found that in MLB, the visibility and popularity from an MVP season could lead to a more loyal fan base and sustained interest in the team, cultivating higher long-term attendance and merchandise sales.[6] This prolonged

DOI: 10.36838/v7i8.6

fan engagement is typically enhanced by strategic marketing and promotional efforts, including premium seating, special events, and merchandise linked to the MVP player, which serve to maximize both immediate and long-term financial benefits.

Contemporary research further bolsters the notion that star players significantly influence game attendance and financial outcomes beyond their direct contributions to team success. Humphrey and Johnson concluded that superstar players generate substantial externalities, increasing both home and away game attendance. Their findings specifically suggest prominent basketball players such as Michael Jordan attracted over 5,000 additional fans per game, emphasizing the idea that the presence of a marquee player can have widespread economic implications for a sports franchise.[7]

Moreover, Slusser deduces that the presence of achieving players is positively correlated with increased attendance, indicating that fans are more likely to attend games featuring high-performing players. These results underscore the significant role star players play in driving fan engagement for franchises.[8]

In conclusion, the literature strongly supports the notion that MVP seasons and star player performances have a significant positive impact on teams' financial outcomes. This is particularly evident in leagues like MLB and the NBA, where the opportunity to witness exceptional athletic performances drives fan engagement, ticket sales, and brand loyalty. Thus, teams can strategically leverage star players to enhance both immediate and future profitability.

Methods

This study examines the financial factors influencing Major League Baseball (MLB) teams, focusing on two key dependent variables: gate receipts and attendance. The data spans from 2009 to 2019, with each MLB team represented as an individual in the panel, labeled from 1 to 30.

Data Collection:

Dependent Variables:

Gate Receipts: The revenue generated from ticket sales for each team.

Attendance: The number of attendees at each game for each team.

Quantitative Variables:

MVPYear1: Indicates if a player from the team won the MVP award in the previous year.

MVPYear2: Indicates if a player from the team won the MVP award two years prior.

MVPYear3: Indicates if a player from the team won the MVP award three years prior.

Record Wins: The number of wins recorded by the team in a season.

Playoff Appearances: Indicates if the team made the playoffs.

Division Win: Indicates if the team won its division.

This study estimates the following panel regression model with year and team fixed effects:

Equation 1:

$$Perf_{i,t}=\beta_1 MVP1_{i,t}+\beta_2 MVP2_{i,t}+\beta_3 MVP3_{i,t}+\partial X_{i,t}+j_t+\rho_i+\varepsilon_{i,t}$$

Perf is the performance of the team gate receipts or attendance for team (i) in year (t)

MVP is an indicator variable for teams (i) that had an MVP one year ago (MVP1), two years ago (MVP2), or three years ago (MVP3), in year (t)

X is a vector of control variables for team (i) in year (t) that include the team's number of wins that year, an indicator for whether the team won the division, and an indicator for whether the team made the playoffs.

j are fixed effects

p are fixed effects

A panel regression is well-suited for this study as it accounts for cross-sectional (team-specific) and time-series (year-specific) variations, providing robust insights into the dynamic nature of financial outcomes across different teams over time.

Result and Discussion

Analysis of Gate Receipts:

Summary Statistics:

The mean value of gate receipts is 83.739 million with a standard deviation of 56.220 million, reflecting variability across teams. This large standard deviation suggests significant differences in ticket revenue among the MLB teams, influenced by factors like team performance, fan base size, and stadium capacity.

Goodness of Fit:

R-squared: 0.294, indicating that 29.4% of the variation in gate receipts is explained by the model. This suggests that while the model accounts for some of the variation, a significant portion remains unexplained, pointing to other potential factors not included in the model. Adjusted R-squared: 0.210, which adjusts for the number of predictors in the model. This indicates that after accounting for the number of variables, about 21% of the variation in gate receipts is explained. The lower adjusted R^2 compared to R^2 suggests that some predictors may not add substantial explanatory power.

Significance Tests:

F-statistic: 20.420 with a p-value of <0.0001, suggesting that the model is statistically significant. The p-value indicates a less than 0.01% chance that the relationships observed are due to random variation, confirming that the combined effects of the predictors meaningfully influence gate receipts.

Coefficients:

MVPYear1: Positive effect on gate receipts (estimate: 8.132, p-value: 0.055). This indicates that a team's MVP win in the current year has a marginal impact on increasing gate receipts, though it is just above the threshold of statistical significance (0.05). This suggests that the effect could be relevant but warrants further investigation.

MVPYear2: Positive effect (estimate: 15.804, p-value: 0.000), indicating that an MVP win from the previous year significantly boosts gate receipts. The low p-value highlights a

strong relationship, as it suggests the impact is unlikely to be due to chance.

MVPYear3: Strong positive effect (estimate: 19.565, p-value: <0.0001), emphasizing that an MVP win from two years ago has a lasting positive effect on ticket sales. This suggests that MVP recognition can continue to attract fans well beyond the season in which the award is won.

Record Wins: Positive impact (estimate: 0.448, p-value: 0.014). This means that as a team wins more games, their gate receipts tend to increase, likely due to heightened fan interest and demand for tickets. Playoff Appearances: Not statistically significant (p-value: 0.401), indicating that merely making the playoffs does not have a significant direct impact on gate receipts.

Division Win: Significant positive impact (estimate: 10.128, p-value: 0.007). Winning a division correlates with higher gate receipts, likely reflecting the prestige and increased fan engagement associated with this achievement.

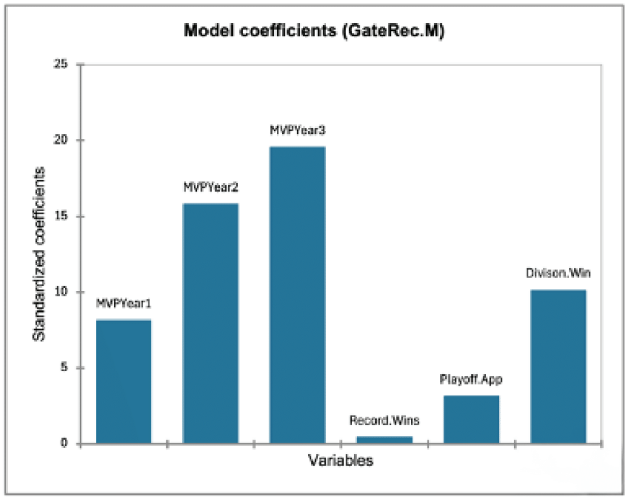


Figure 1: This figure illustrates the estimated coefficients from the regression model evaluating the impact of MVP seasons and other performance variables on gate receipts. It highlights the magnitude and statistical significance of each predictor, demonstrating that MVP seasons, particularly from prior years, and team success metrics, such as division wins, positively influence ticket revenue. This figure highlights that MVP seasons, particularly from two and three years prior, significantly boost revenue, with MVPYear3 having the strongest long-term effect.

The mean gate receipts of $83.739 million, with a substantial standard deviation of $56.220 million, as shown in 1, underscore the variability across MLB teams, influenced by factors such as fan base size, team performance, and market dynamics. Key findings reveal a notable positive influence of MVP seasons, particularly in MVPYear3, which demonstrates a lasting financial boost with a significant coefficient of 19.565 and a p-value of <0.0001, as showcased in Figure 1. This impact suggests that fan excitement and engagement remain elevated long after the MVP season.

Table 1: This table presents the estimated coefficients, standard errors, and significance levels for the regression model predicting gate receipts. Key findings include the positive and significant effects of MVP wins from prior years and record wins, with division wins also contributing to increased revenue. The model captures the relationship between individual achievements, team performance, and financial outcomes. The findings reveal that MVP wins from prior years have a sustained positive impact, with MVPYear3 showcasing the strongest correlation to increased ticket revenue.

	Estimate	Std. Error	t-value	Pr(>|t|)
MVPYear1	8.132	4.223	1.926	0.055
MVPYear2	15.804	4.414	3.581	0.000
MVPYear3	19.565	4.702	4.162	<0.0001
Record.Wins	0.448	0.129	3.459	0.001
Playoff.App	3.142	3.735	0.841	0.401
Divison.Win	10.128	3.754	2.698	0.007

Further, as shown in Table 1, there is a significant correlation between team success and gate receipts. The positive coefficients for record wins (0.448) and division wins (10.128) indicate that higher win rates are associated with increased revenue from gate receipts.

Additionally, the R-squared value of 0.294 and adjusted R-squared value of 0.210 indicate that, while the model captures some of the variations in gate receipts, additional unexplored factors likely contribute to revenue differences.

Analysis of Attendance

Summary Statistics:

The average attendance per team is 29,973.221 with a standard deviation of 8,115.263, indicating variability across teams. The substantial standard deviation reflects differing fan bases and stadium capacities, suggesting that factors like team performance, location, and market size have notable impacts on attendance.

Goodness of Fit:

R-squared: 0.341, indicating that 34.1% of the variation in attendance is explained by the model. This means that the chosen variables, such as MVP status and record wins, provide a reasonable explanation of attendance changes, though other factors (like economic conditions or promotional activities) might also play a role. Adjusted R-squared: 0.262, accounting for the number of predictors. The adjusted R^2 being lower than R^2 suggests that while the model explains some variance, the predictive power is somewhat diminished when adjusting for the number of independent variables.

Significance Tests:

F-statistic: 25.318 with a p-value of <0.0001, indicating the model's overall significance. The very low p-value implies that the relationship between the predictors and attendance is unlikely to be due to random chance.

Coefficients:

MVPYear1: Significant positive effect on attendance (estimate: 2308.469, p-value: 0.005). This means that an MVP win in the current year substantially increases game attendance, likely due to heightened interest from fans wanting to see the star player.

DOI: 10.36838/v7i8.6

MVPYear2: Significant positive effect (estimate: 2607.663, p-value: 0.003), showing that an MVP win from the previous year continues to draw fans to games. This suggests a lasting influence of star power on fan engagement.

MVPYear3: Significant positive effect (estimate: 2461.880, p-value: 0.008). Even two years after an MVP win, the award's impact on attracting fans persists, reflecting the importance of star players in maintaining high attendance.

Record Wins: Strong positive relationship (estimate: 165.470, p-value: <0.0001), indicating that a better win record directly translates to higher attendance. This is likely because fans are more interested in watching a winning team.

Playoff Appearances: Not significant (p-value: 0.981), suggesting that making the playoffs does not have a direct impact on regular season attendance.

Division Win: Not statistically significant (p-value: 0.181), indicating that a division win does not necessarily boost regular season attendance figures.

Table 2: This table summarizes attendance-related data, showcasing descriptive statistics for attendance and predictors such as MVP years, record wins, and postseason achievements. The data indicates that MVP seasons contribute to higher average attendance over multiple years.

Variable	Observations	Obs. with missing data	Obs. without missing data	Minimum	Maximum	Mean	Std. deviation
Attendance	330	0	330	10013.0	49066.000	29973.221	8115.263
MVPYear1	330	0	330	0.000	1.000	0.061	0.239
MVPYear2	330	0	330	0.000	1.000	0.055	0.227
MVPYear3	330	0	330	0.000	1.000	0.048	0.215
Record.Wins	330	0	330	47.000	108.000	80.988	11.847
Playoff.App	330	0	330	0.000	1.000	0.303	0.460
Divison.Win	330	0	330	0.000	1.000	0.200	0.401

The analysis of attendance data provides valuable insights into the factors driving fan engagement and attendance at MLB games. The average attendance across all teams is 29,973 per game, with a significant standard deviation of 8,115, as noted in Table 2. This variability reflects differences in stadium capacities, market sizes, and fanbase engagement across teams. Table 4 highlights the strong positive impact of MVP seasons, with the coefficients for MVPYear1 (2,308.469, p-value 0.005), MVPYear2 (2,607.663, p-value 0.003), and MVPYear3 (2,461.880, p-value 0.008) all indicating statistically significant contributions to game attendance. These findings demonstrate that the excitement surrounding an MVP-caliber player extends well beyond the award year, maintaining high levels of fan interest for multiple seasons.

Table 3: This table details the coefficients from the attendance regression model, emphasizing the significant positive impact of MVP wins on fan turnout. It also highlights the importance of record wins in driving attendance, while other variables like playoff appearances and division wins show weaker or insignificant effects on regular-season attendance figures. The results show MVP wins boost attendance not just in the award year but also in subsequent years, reinforcing the impact of star players.

	Estimate	Std. Error	t-value	Pr(>|t|)
MVPYear1	2308.469	821.492	2.810	0.005
MVPYear2	2607.663	858.672	3.037	0.003
MVPYear3	2461.880	914.656	2.692	0.008
Record.Wins	165.470	25.167	6.575	<0.0001
Playoff.App	-17.333	726.645	-0.024	0.981
Divison.Win	979.872	730.320	1.342	0.181

Table 3 also reveals a strong relationship between team performance and attendance, as shown by the significant coefficient for record wins (165.470, p-value <0.0001), confirming that fans are drawn to winning teams. However, other metrics, such as playoff appearances (p-value 0.981) and division wins (p-value 0.181), do not exhibit significant effects, suggesting that regular-season success and individual player performance may be stronger drivers of attendance than postseason outcomes. The model's R-squared value of 0.341 and adjusted R-squared of 0.262 indicate that 34.1% of the variation in attendance is explained by the included variables, though other factors—such as economic conditions, promotions, or ticket pricing strategies—may also influence attendance.

Conclusion

This research affirms the positive impact that MVP seasons have on team financial outcomes, particularly gate receipts and attendance. Teams with an MVP player experience a tangible financial boost, which compounds over successive MVP-caliber seasons. Teams with an MVP player enjoyed a tangible financial boost, with gate receipts increasing by approximately $8.1 million in the award year and up to $19.6 million three years post-award. Surprisingly, the impact of an MVP season is not most pronounced in the award year itself but rather reaches its zenith in the third subsequent year, with a $19.6 million increase in gate receipts. The lagged peak contradicts assumptions of temporary excitement and points toward a persistent effect on fan interest. However, the relatively low R-squared values (0.294 for gate receipts and 0.341 for attendance) imply that while MVP seasons and performance metrics explicate a portion of the variance, other variables also play a role in determining financial outcomes.

Moreover, the excitement generated by an MVP player can have a lasting impact beyond the immediate increase in attendance and revenue. Higher fan turnout can lead to enhanced merchandising opportunities, greater media exposure, and a stronger brand presence for the team. These findings align strongly with Rosen's "Superstar Effect," assuring that exceptional individual performance drives revenue not only in the award-winning season but over a multi-year period. Future research could build on these findings by exploring several additional factors. One area worth investigating is the long-term financial effects of MVP seasons, such as how they influence sponsorship deals, media contracts, and overall team valuation over multiple years. Furthermore, studies could investigate the role of fan loyalty and demographics to see how MVP performances drive long-term fan engagement and whether this translates into sustained financial growth. To the administrative side of team operations, this implies that capturing momentum with MVP seasons using strategic marketing, flexible ticketing, and better merchandise may help sustain and accelerate financial returns. As well as increasing investments into longer-term branding by players for even greater leveraging of fan devotion and involvement. A potential confounding factor is that MVPs are more likely to play for teams in larger markets or already successful franchises, which may inherently have higher atten-

dance and revenue levels. This makes it difficult to fully isolate the MVP effect from the market and team context.

Additionally, examining the interplay between external economic conditions and the financial benefits of an MVP season could provide more context for understanding how different markets react to star performances. Comparative research across teams in large and small markets or studies analyzing the effects of MVP seasons during periods of economic downturn or growth might yield valuable insights. By broadening the scope of analysis, future research could provide a more holistic understanding of how MVP seasons affect MLB teams, offering valuable insights for team management, stakeholders, and sports economists looking to maximize the economic potential of superstar players.

Limitations

The study acknowledges several limitations that may affect the accuracy and generalizability of its findings. First, it is constrained by the time range of data (2009-2019), which means that it does not account for longer-term trends or more recent changes in Major League Baseball (MLB) that could

influence how MVP seasons impact financial metrics. For example, shifts in league-wide attendance trends or revenue structures in the post-2019 era could yield different results.

Another limitation is the exclusion of broader economic conditions, such as local or national economic downturns, which could significantly impact fan spending behavior. During recessions, for instance, consumers are more likely to be more frugal with spending, including entertainment expenses like attending baseball games. This means that even an MVP season may not generate the same level of financial boost in a volatile economy as it would in a more prosperous economy. The study also does not fully account for market size variations. Teams in larger, wealthier markets may experience more substantial revenue increases due to higher baseline attendance and stronger local economies, compared to teams in smaller markets where fan bases and disposable incomes are smaller. This disparity could skew the financial impact of an MVP season, as teams in smaller markets may not see the same level of increase in gate receipts and attendance.

Another key factor that was not deeply explored is ticket pricing strategies. Teams often adjust ticket prices during or following MVP seasons to capitalize on increased demand. However, if prices rise too steeply, it could limit attendance growth, particularly for lower-income fans. Conversely, teams that do not increase ticket prices could miss out on potential revenue from premium pricing. The balance between demand elasticity and pricing strategies is crucial but not thoroughly examined in this study.

Finally, the study doesn't fully address external competition from other sports or entertainment options, which may affect attendance and revenue. In markets where multiple professional sports teams or large-scale entertainment events compete for consumer attention, the financial boost of an MVP season could be diluted.

By addressing these limitations in future research, a more comprehensive understanding of the financial impact of MVP seasons can be achieved, providing clearer insights into how external and internal factors shape revenue outcomes across different MLB teams and seasons.

Acknowledgments

I want to thank Professor Jason Damm of the University of Miami for his invaluable mentorship throughout the creation of this paper.

References

1. Rosen, S. The Economics of Superstars. *Am. Econ. Rev.* **1981**, *71* (5), 845–858. http://www.jstor.org/stable/1803469.
2. Torres-Reyna, O. Panel Data Analysis: Fixed and Random Effects using Stata. *Princeton University*, **2007**. https://www.princeton.edu/~otorres/Panel101.pdf.
3. Hausman, J. A.; Leonard, G. K. Superstars in the National Basketball Association: Economic Value and Policy. *J. Labor Econ*. **1997**, *15* (4), 586–624. https://doi.org/10.1086/209839.
4. Horowitz, I. Review of *The Business of Major League Baseball*, by G. W. Scully. *Rev. Ind. Organ*. **1990**, *5* (3), 115–118. http://www.jstor.org/stable/41798319.
5. Berri, D.; Schmidt, M.; Brook, S. Stars at the Gate: The Impact of Star Power on NBA Gate Revenues. *Res. Gate*. **2004**. https://www.researchgate.net/publication/247739515_Stars_at_the_Gate_The_Impact_of_Star_P ower_on_NBA_Gate_Revenues.
6. West Virginia University. An Analysis of Major League Baseball Attendance with Superstar Player Performance and Team Quality Metrics. *WVU Res. Repo*s. [n.d.] https://researchrepository.wvu.edu/econ_working-papers/29/.
7. Humphreys, B. R.; Johnson, C. The Effect of Superstar Players on Game Attendance: Evidence from the NBA. *SSRN Scholarly Paper*, **2017**. https://ssrn.com/abstract=3004137.
8. Slusser, A. The Star Player Effect: Does the Quality of a Player Impact Attendance in the NBA? *Honors Thesis*, Ball State University, **2021**. https://cardinalscholar.bsu.edu/handle/123456789/2021.

Author

Michael Bronshteyn is a junior at Calabasas High School in Calabasas, California. He is interested in analyzing culture by utilizing economic frameworks

DOI: 10.36838/v7i8.6

IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

An Alternative Composite Score Using Easier to Access Data to Determine the Probability of Recurrence in HER2-ve Breast Cancer

Rishi V. Pai

Northview High School, 10625 Parsons Rd., Johns Creek, Georgia, 30097, USA; pai.rishi0709@gmail.com

ABSTRACT: Luminal breast cancers (BC) are the most common subtype, with human epidermal growth factor receptor 2 negative being the most prevalent. Currently, the genetic-based Oncotype DX score is the widely used metric for determining recurrence probability, but the 21-genes it utilizes are difficult and time-consuming to obtain data for. This study aimed to use machine learning classification to predict composite scores for recurrence probability based on easier-to-access data, selected from feature importance identified by the model(s). Four classification models were trained and tested to predict BC recurrence: random forest (RF), logistic regression (LR), gradient boosting (GBM), and decision tree (DT). The models were compared by their F1 score. The most significant variables from the best-performing model trained a Calibrated GBM (chosen as the final predictor as it had the highest F1 score at 0.24) to predict composite scores from 0-10. The predictive model returned scores with ~92% accuracy. Findings showed that the GBM is not reliable for solely binary recurrence prediction due to poor performance metrics, but it serves as a promising tool for predicting composite scores for recurrence probability. These scores could be utilized in clinical settings to determine intervention plans to improve patient prognosis.

KEYWORDS: Robotics and Intelligent Machines, Machine Learning, Predictive Analytics, Breast Cancer, Prognosis.

Introduction

Luminal breast cancers (BC) account for approximately 65% of all BC cases, being the most common subtype.[1, 2] Human epidermal growth factor receptor 2 negative (HER2-negative) is the most common for early stage BC, as represented in Figure 1.[3, 4] The number of deaths from BC is estimated to rise to 2.9 million annually.[5] Thus, early diagnosis and effective treatment are crucial for prognosis efforts.[1, 6]

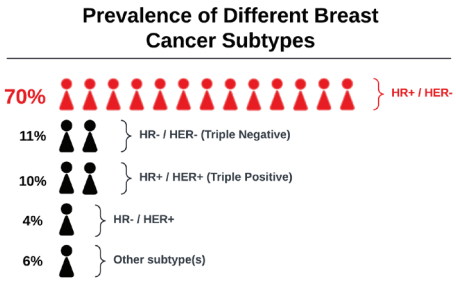


Figure 1: Prevalence of different BC subtypes. Data gathered from the National Cancer Institute SEER (Surveillance, Epidemiology, and End Results Program).[4] The visual presents the HER2-negative subtype as the most prevalent of all breast cancer cases (~70%).

Machine learning's (ML) ability to analyze large amounts of data makes it promising as a tool for predicting BC recurrence.[7, 9] Some known important factors include histological grade, tumor size, nodes, genomic score, and the Ki67 proliferation index.[3] Thus far, the recurrence of BC and its factors have not been thoroughly studied through machine learning techniques as a consequence of patient recurrence data rarely being available in accessible datasets.[10]

The widely used metric score for BC recurrence probability is the Oncotype DX score, a 21-gene recurrence tool that examines the activity of genes in the breast tumor tissue.[2, 11] The scores range from 0 to 100, with higher scores reflecting a higher probability of BC recurrence as well as the likelihood of benefitting from chemotherapy and hormonal therapy. Patients who have Oncotype DX scores above 26 benefit most from chemotherapy.[2] The score was developed mainly for estrogen receptor-positive (ER-positive) and HER2-negative BC, and it gives an accurate estimation of recurrence probability.[11]

Genetic patient data for the Oncotype DX score is generally difficult, tedious, or time-consuming to obtain, so an equally accurate recurrence probability score calculated from easier-to-access data would be more efficient. This study aims to develop a machine learning classification model to predict the recurrence of BC based on clinical, histological, immunohistochemical, molecular biology, and treatment patient data. A second classifier trained by data selected from feature importance identified by the original model(s) will be tested to predict accurate composite scores from 0 to 10 for recurrence probability.

Methods

All computations were performed on Python version 3.11.7 on JupyterLab 4.0.11. All machine learning models and techniques were provided by the *scikit-learn* package in Python. The research methods are outlined in Figure 2.

DOI: 10.36838/v7i8.7

Data Collection and Curation:

A publicly available and anonymous Oncotype DX patient dataset containing 321 patients was originally obtained from the Georges Francois Lecler Cancer Centre and North Trévenans County Hospital, both in France.[12] The dataset was accessed via data.world, which provided a convenient platform for data retrieval. The dataset was originally in French and was translated into English.

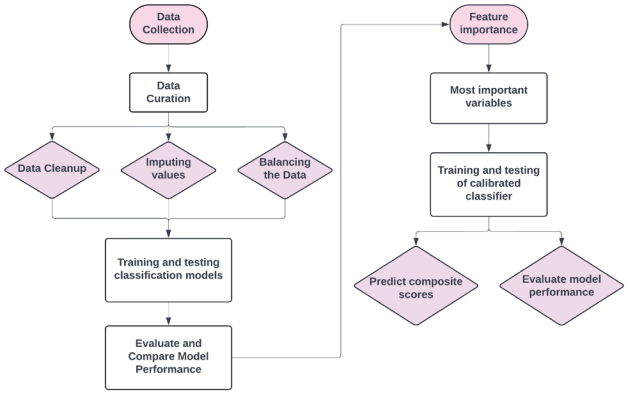


Figure 2: Workflow of the research methods. The chart displays the process from sole BC recurrence prediction to the final composite score prediction methods.

The columns *NBlock* and *Block ODX* were removed as they had no significance on patient recurrence in the context of the data, since they were present for patient identification. The *oncogenetic consultation, mutation, HER2 IHC,* and *HER2 SISH* columns were removed as they had 70% or more missing information. Imputing data for these columns would not be a proper representation of patient data as there is not enough basis for estimates. The *Stage pTNM* column was removed because it had too many unique values to be significant enough for predicting BC recurrence. While mutation patterns are strong indicators of cancer recurrence, the data set used in this study had a weak basis for clinical staging correlation and would give biased results if the empty data were to be predicted. Finally, the *last contact date* and *last contact status* columns were removed. The date provides no information and does not have a recurrence metric. Pre-testing showed that these columns had little to no impact on changing recurrence predictions, as the most prominent variables consisted of clinical, histological, immunohistochemical, molecular biology, and treatment data native to the dataset. The few deceased patients in the dataset in the *last contact status* column were removed to keep only patients who survived from HER2-negative BC.

Any patients who had blank data for the *Recurrence* column were removed from the dataset. For the remaining data columns that had missing information, these values were filled in using mode imputation to ensure data consistency for model training. Additional data curation included removing secondary or tertiary values from cells that had more than one value and changing all classification columns with string/character data to numeric classification (0, 1, 2, etc)

Patient data for the *node type, histological type,* and *histological subtype* columns were represented with a range of values. These columns were expanded to better visualize the types of each that occurred for each patient. For example, a patient could have either a 0, 1, or 2 for the *node type* column, with each value representing 'no sentinel node or dissection' (column 1), 'sentinel node' (column 2), or 'axillary dissection' (column 3), respectively. The original column was split into three, with each patient recording either a 0 (does not have) or a 1 (does have). A patient who has only sentinel and axillary dissection nodes would have a 0 for column 1, and a 1 for columns 2 and 3.

The dataset was standardized for columns that had numeric values using *scikit-learn* pre-processing. This was done to ensure uniformity and consistency between data columns for model efficiency. After data curation, the dataset was split 65% for model training and 35% for model testing, resulting in 191 training patients and 104 testing patients.

The provided recurrence column had classes 0 and 1, with 0 representing no recurrence of BC and 1 representing recurrence. This column was highly imbalanced, with more than 90% of the patients recording class 0 (non-recurrence) and a much smaller proportion for class 1 (recurrence). To prevent overfitting for the models, the majority class was down-sampled, and the minority class was up-sampled to balance the target column for the training set. This resulted in a training dataset with a combined 92 rows: 50 for class 0 and 42 for class 1.

Model Training and Testing for Binary BC Recurrence Prediction:

Four classification models were used initially to test BC recurrence prediction: random forest (RF), logistic regression (LR), gradient boosting (GBM), and decision tree (DT). These models were chosen for their previous usage in BC studies involving diagnosis and prognosis analysis, and primarily for their ability to record feature importance, which was considered for predicting the composite scores later in the study. Each of the models was evaluated with *F1 score (harmonic mean of precision and recall), accuracy (proportion of total correct predictions), precision (proportion of predicted positive instances that are true positives), recall (sensitivity; proportion of true positive instances correctly identified), and specificity (the proportion of true negative instances correctly identified).* The models were compared by their respective F1 scores. Receiver operating characteristic (ROC) curves and confusion matrices were created for each model to compare efficacy.

Random Forest (RF). The random forest model is a type of supervised learning, where the model uses patterns in the dataset to make predictions based on labeled data.[13] Random forest is an expansion on decision trees;[13, 14] As the method may suggest, random forest models typically perform better than decision tree algorithms, as it is a multi-faceted approach.[14] Decision trees split data into subsets at each node by choosing the feature that best separates the data, and repeat this process recursively until reaching a final prediction.[13] In a random forest, each "tree" votes on a prediction, and the class with the highest number of votes is the final prediction.[13-15] Random forest is typically favored for medical studies and has shown

sufficient results.[14,16] The random forest algorithm is presented in Figure 3.

Logistic Regression (LR). The logistic regression algorithm finds the relationship between features and outcome probability through a sigmoidal curve. Simply, logistic regression returns the likelihood of an outcome when given individual features.[13,17] The model produces single numerical values from 0 to 1 from numeric features.[13]

Gradient Boosting (GBM). A gradient boosting classifier sequentially builds an ensemble of weaker models, typically decision trees, where each new model is trained to correct errors from the previous ones;[18] Gradient boosting gradually improves overall predictive accuracy by fitting consecutive models.[18] Boosting is an improvement on simpler ensemble techniques like decision trees or random forests by iteratively training new models from prior mistakes.

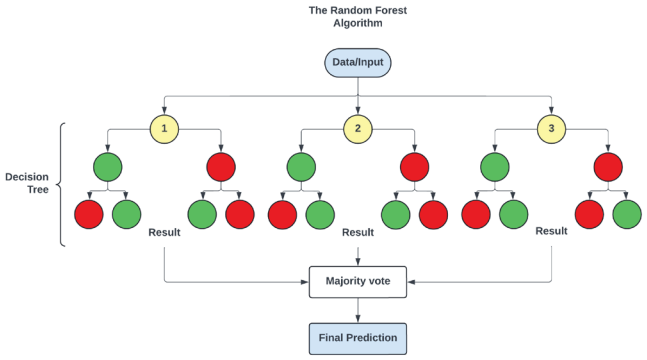


Figure 3: Visual of the random forest algorithm. The algorithm is an expansion of decision tree classification, seen in the diagram as a culmination of several subtrees. The visual presents how the algorithm uses "votes" from each subtree to reach a conclusion.

Each model is able to return the variables that had the most impact in predicting the target column (feature importance). It can be thought of as determining the "decision-making power" for each data variable.

Predicting the Recurrence Probability Composite Scores:

The same 92-row training dataset for binary recurrence prediction was used for training and testing the model to generate composite scores, with an 80:20 training/testing split. However, the model was trained only with the top 21 features, or the features that had a significant numeric score. Significant numeric scores were regarded as scores that had a positive value, meaning that they contributed to a notable percentage of the prediction. For example, a score with a value of 0.15 represents a 15% contribution to the total importance of the recurrence/non-recurrence prediction. The classifier for composite score prediction was trained with the same algorithm(s) as those of the model for binary BC prediction. While the variable range was limited in this regard, through feature importance selection, the model still utilized the provided data to identify patterns between patient data and their assigned Oncotype DX score.

Out of the top 21, variables that had no real impact on patient prognosis were removed, such as diagnosis year and birth year. The model that was used for score prediction was a calibrated GBM, as it had the highest F1 score compared to the other classifiers when predicting recurrence. *** Explained in Results.**

Regular classifiers will return raw scores while a calibrated one adjusts scores to provide more accurate probabilities, meaning that the predictions more accurately reflect the true likelihood of an outcome. Calibration techniques provide a more proper reflection of recurrence probability through the composite scores. The recurrence target column remained in the training set but was removed in the testing set for the model to predict scores purely from patient data.

Result and Discussion

All charts/graphs were created using the matplotlib package in Python.

Evaluation of the Models for Recurrence Prediction:

The five recorded performance metrics for each model are presented in Table 1. In this context, classes are predicted by the model (true positives - recurrence and true negatives – non-recurrence), so the F1 score is a preferred metric rather than accuracy to compare model performance, as it is the numeric mean of precision and recall.

Table 1: Performance metrics of the four classifiers to predict BC recurrence. The table shows that the Gradient Boosting (GBM) classifier resulted in the highest F1 Score (indicated in bold), but did not perform sufficiently in other metrics.

Model	F1 Score	Accuracy	Precision	Recall	Specificity
Random Forest	0.12	0.89	0.25	0.11	0.97
Logistic Regression	0.19	0.75	0.13	0.33	0.79
Gradient Boosting *	**0.24**	0.75	0.16	0.44	0.78
Decision Tree	0.08	0.76	0.06	0.11	0.84

Apart from accuracy and specificity, all models returned relatively low performance metrics (values closer to 1.0 indicate better performance in the field). Random forest showed the highest accuracy (0.89), the highest specificity (0.97), and the highest precision (0.25) compared to the other models, but had a low F1 score (0.12). As briefly mentioned in the methods, the Gradient Boosting Classifier was chosen as the final recurrence predictor as it had the highest F1 score (0.24).

The ROC curve was plotted for the Gradient Boosting Classifier, as shown in Figure 4. The curve shows model performance by plotting the true positive rate (y-axis) compared to the false positive rate (x-axis).

DOI: 10.36838/v7i8.7

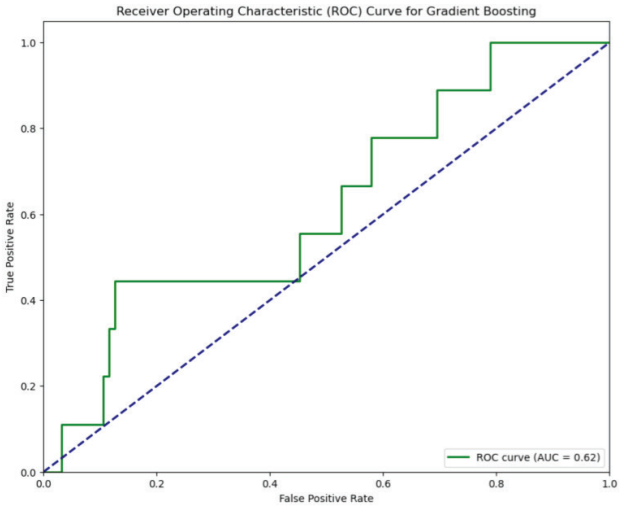


Figure 4: The ROC curve for the GBM compared to the area under the curve (AUC). The figure shows that the model's correlation between false and true positive predictions is similar to that of random guessing, indicated by its close distance to the AUC line.

The Area under the curve (AUC; blue-dotted line) is representative of random guessing. Ideally, peak model performance would result in an ROC curve that is closest to the top left corner of the graph, somewhat like a logarithmic function. Figure 4 shows that although the ROC curve for the model is higher than the AUC curve, it does not necessarily achieve high performance. It reflects only a slightly better performance than random guessing.

The confusion matrix is another method to evaluate the performance of a classification model by displaying the number of true/false positives and negatives predicted by the model for the testing dataset. The confusion matrix for the Gradient Boosting Classifier is outlined in Table 2.

Table 2: Confusion matrix for the testing dataset for the Gradient Boosting Classifier. The table shows strong results for the GBM classifier's negative prediction capability. Conversely, the GBM classifier is unreliable for predicting positive instances.

		Predicted	
		No (0)	Yes (1)
Actual	No (0)	True Negative **74**	False Positive **21**
	Yes (1)	False Negative **5**	True Positive **4**

The testing dataset consisted of 104 rows but was imbalanced for the recurrence column due to the original Oncotype DX dataset, so there was a considerably higher number of negative predictions. Out of 79 cases of negative instances (non-recurrence), the model was able to predict 74 accurately, with the other 5 predictions returning false negatives. However, out of 25 that did have recurrence, the model was only able to predict 4 accurately, with the other 21 returning false positives. These results, displayed in the confusion matrix, show that the model is accurate at predicting negative instances (~94% correct predictions) but poor at predicting positive instances (16% correct predictions). These metrics correlate to the model's low scores for specificity (true negative rate) and sensitivity (true positive rate) from Table 1. Thus, the model is highly accurate for predicting non-recurrence in patients but is not very accurate at predicting recurrence.

Composite Score Results:

The feature importance for the GBM is presented in Figure 5.

The five most important features for recurrence prediction identified by the model were **diagnosis year, surgery, oncotype score, progesterone receptor (PR) Allred score, and PR percentage,** with numeric scores of ~0.25, ~0.17, ~0.10, and ~0.06, respectively.

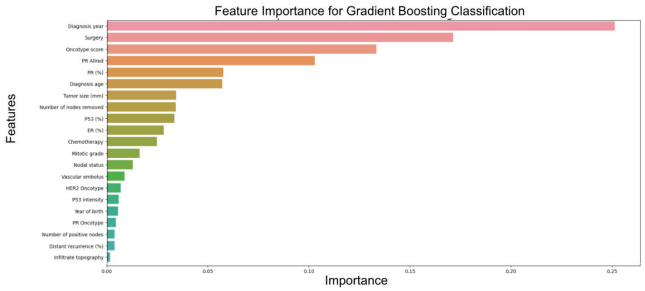


Figure 5: A bar graph that visualizes and compares the feature importances for the Gradient Boosting Classifier, showing which data columns (variables) are considered the most important for predicting BC recurrence. Features that did not have significant values were excluded from the analysis.

Examples of composite scores for patients in the testing dataset are presented in Table 3. The calibrated Gradient Boosting Classifier was able to accurately predict composite scores to reflect the probability of BC recurrence (~92% accuracy). A score greater than 5 and closer to 10 indicates a higher probability of recurrence. Conversely, a score less than 5 and closer to 0 indicates a lower probability of recurrence, a similar system to Oncotype DX. The calibrated model was trained with all the top 21 variables from the feature importance graph (Figure 5), excluding *diagnosis year, oncotype score,* and *year of birth.* Diagnosis year and year of birth were insignificant variables as the "diagnosis age" was already provided for each patient. Specifically, the diagnosis year and year of birth columns were values that had little correlation to the recurrence predictions. The distant recurrence percentage was included as a variable in composite score training to serve as a proxy for metastatic potential. The metric reflects the likelihood of the tumor spreading to distant organs over time and helps the training model better understand the biological behavior and aggressiveness of the tumor. The distant recurrence percentage is

linked to molecular biology, which allows the model to base predictions regarding metastatic risk.

Table 3: A sample of 10 patients from the testing dataset with new predicted recurrence probability scores. The scores were rounded to the nearest two decimal places. The composite scores were predicted with ~92% accuracy to the corresponding original Oncotype-DX score for the patient.

Patient	New Recurrence Score
1	3.48
2	9.50
3	8.66
4	1.22
5	1.56
6	9.80
7	10.0
8	2.91
9	9.14
10	4.64

Discussion:

This study aimed to analyze prognostic factors for BC recurrence and use those factors to develop a model to predict composite scores to reflect recurrence probability as an alternative to Oncotype DX.

Out of the four classification models that were tested (random forest, logistic regression, gradient boosting, and decision tree), the Gradient Boosting Classifier was chosen as the final model as it had the highest F1 score (0.24). This model had a relatively high overall accuracy (0.75) for predicting BC recurrence. It is important to note that the model's performance in predicting positive instances (having recurrence) is quite poor, as indicated by the low precision and recall (0.16, and 0.44, respectively), as well as the disparities of the confusion matrix (Table 2) true/false positive prediction for the testing dataset. The high specificity of the model (0.78) and the confusion matrix again show that the model is sufficient at identifying negative instances. The discrepancy between positive and negative case predictions is likely due to the imbalance of the target column in the dataset.

The top 21 features of highest importance in descending order include: *diagnosis year, surgery, oncotype score, PR Allred, PR %, diagnosis age, tumor size (mm), number of nodes removed, P53 %, ER (estrogen receptor) %, chemotherapy, mitotic grade, nodal status, vascular embolus, HER2 oncotype, P53 intensity, year of birth, PR oncotype, number of positive nodes, distant recurrence %, and infiltrate topography*. These factors include clinical, histological, immunohistochemical, molecular biology, and treatment data. The results fairly align with previous BC prognostic studies that showed tumor size as one of the most important variables for recurrence prediction.[3]

While the Gradient Boosting Classifier might not be a reliable tool for solely binary prediction of BC recurrence, it appears as a promising method for predicting the probability scores of recurrence (~92% accuracy). The GBM was inconsistent with recurrence/non-recurrence predictions in the first step of the methodology, but proved sufficient in its ability to correlate data points to previously calculated Oncotype DX scores. This evaluation allowed it to identify what data constitutes a specific score, ensuring accuracy in the newly generated composite scores. The predicted scores are a fairly accurate alternative to the widely used Oncotype DX score, as they utilize more easily accessible patient data for calculating recurrence probability rather than the 21-genes used for calculating the Oncotype DX score.

■ Conclusion

In this study, four classification machine learning models were trained and tested with recurrence data from HER2-negative BC patients in France to determine the most important factors for recurrence and use them to predict an efficient composite score based on easier-to-obtain data to reflect a patient's chances of BC recurrence.

The Gradient Boosting Classifier overall is not a great method for predicting BC recurrence in hospital settings. While not sufficient for solely recurrence prediction, the GBM is practical for predicting composite scores. However, the model does show promise, but requires refining for recurrence prediction. This predictive score could be utilized in hospital settings to determine correct intervention plans to improve patient prognosis.

Future Work:

The Oncotype DX dataset used in this study was fairly limited with the number of patients, but the key downside was the imbalance of the recurrence data. More than 90% of the patients recorded non-recurrence, while a few recorded recurrences, which required artificial balancing of the data. While still an accurate estimation of recurrence data, this placed a bias on negative instance prediction compared to positive instance prediction for the testing datasets. Training and testing a model with a larger BC recurrence dataset with a larger scope, with more true and balanced prognosis data, should be done to yield higher performance results for classification models.

Additionally, common clinical practice is to use one dataset for model training and a separate test for model validation. Regarding breast cancer recurrence, there is a scarcity of publicly available, anonymous datasets with true recurrence data, which is the sole reason that breast cancer recurrence is an under-studied field of oncology. To compensate, the dataset used in this study was split to create separate sets for training and testing, with the testing dataset having no predicted/imputed data to mitigate bias for model validation. Thus, using a larger and more diverse data set is a prospective endeavor to boost performance for the models used in this study.

Applying the methodology and composite score prediction techniques used in this study for other cancers and diseases, such as cardiovascular disease, may help to better understand recurrence probability predictors.

■ Acknowledgments

I would like to thank Jagdeep Podichetty, PhD, at the Critical Path Institute for his guidance and support throughout the research process.

■ References

1. Gilchrist, J. Current Management and Future Perspectives of Hormone Receptor-Positive HER2-Negative Advanced Breast

DOI: 10.36838/v7i8.7

Cancer. *Seminars in Oncology Nursing*. **2024**, *40* (1). DOI: 10.1016/j.soncn.2023.151547
2. Durrani, S.; Al-Mushawa, F.; Heena, H.; Wani, T.; Al-Qahtani, A. Relationship of Oncotype Dx score with tumor grade, size, nodal status, proliferative marker Ki67 and Nottingham Prognostic Index in early breast cancer tumors in Saudi Population. *Annals of Diagnostic Pathology*. **2021**, *51*. DOI: 10.1016/j.anndiagpath.2020.151674
3. Zambelli, A.; Gallerani, E.; Garrone, O.; Pedersini, R.; Rota Caremoli, E.; Sagrada, P.; Sala, E.; Cazzaniga, ME. Working tables on Hormone Receptor Positive (HR+), Human Epidermal growth factor Receptor 2 negative (HER2-) early stage breast cancer: Defining high risk of recurrence. *Critical Reviews in Oncology/Hematology*. **2023**, *191*. DOI: 10.1016/j.critrevonc.2023.104104
4. NIH SEER. *Cancer Stat Facts: Female Breast Cancer Subtypes*. https://seer.cancer.gov/statfacts/html/breast-subtypes.html (accessed 2024-07-20)
5. Sharma, A.; Goyal, D.; Mohana, R. An ensemble learning-based framework for breast cancer prediction. *Decision Analytics Journal*. **2024**, *10*. DOI: 10.1016/j.dajour.2023.100372
6. Zuo, D.; Yang, L.; Jin, Y.; Qi, H.; Liu, Y.; Ren, L. Machine learning-based models for the prediction of breast cancer recurrence risk. *BMC Medical Informatics and Decision Making*. **2023**. DOI: 10.1186/s12911-023-02377-z
7. González-Casto, L.; Chávez, M.; Duflot, P.; Bleret, V.; Martin, A. G.; Zobel, M.; Nateqi, J.; Lin, S.; Pazos-Arias, J. J.; Del Fiol, G.; López-Nores, M. Machine Learning Algorithms to Predict Breast Cancer Recurrence Using Structured and Unstructured Sources from Electronic Health Records. *MDPI*. **2023**, *15* (10). DOI: 10.3390/cancers15102741
8. Liu, Y.; Fu, Y.; Peng, Y.; Ming, J. Clinical decision support tool for breast cancer recurrence prediction using SHAP value in cooperative game theory. *Heliyon*. **2024**, *10* (2). DOI: 10.1016/j.heliyon.2024.e24876
9. Jin, Y.; Lan, A.; Dai, Y.; Jiang, L.; Li, S. Development and testing of a random forest-based machine learning model for predicting events among breast cancer patients with a poor response to neoadjuvant chemotherapy. *European Journal of Medical Research*. **2023**, *28* (394). DOI: 10.1186/s40001-023-01361-7
10. Henriquez Abreu, P.; Seoane Santos, M.; Henriques Abreu, M.; Andrade, B.; Castro Silva, D. Predicting Breast Cancer Recurrence Using Machine Learning Techniques: A Systematic Review. *ACM Computing Surveys*. **2016**, *49* (3), 1-40. DOI: 10.1145/2988544
11. Loggie, J.; Barnes, P. J.; Carter, M. D.; Rayson, D.; Bethune, G. C. Is Oncotype DX testing informative for breast cancers with low ER expression? A retrospective review from a biomarker testing referral center. *The Breast*. **2024**, *75*. DOI: 10.1016/j.breast.2024.103715
12. Zemouri, R.; Omri, N.; Morello, B.; Devalland, C.; Arnould, L.; Zerhouni, N.; Fnaiech, F. Constructive Deep Neural Network for Breast Cancer Diagnosis. *IFAC-PapersOnLine*. **2018**, *51* (27). DOI: 10.1016/j.ifacol.2018.11.660
13. Choi, R. Y.; Coyner, A. S.; Kalapathy-Cramer, J.; Chiang, M. F.; Peter Campbell, J. Introduction to Machine Learning, Neural Networks, and Deep Learning. *Translational Vision Science & Technology*. **2020**, *9* (2). DOI: 10.1167/tvst.9.2.14
14. Minnoor, M.; Baths, V. Diagnosis of Breast Cancer using Random Forests. *Procedia Computer Science*. **2023**, *218*, 429-437. DOI: 10.1016/j.procs.2023.01.025
15. Macauley, B. O.; Aribisala, B. S.; Akande, S. A., Akinnuwesi, B. A.; Olabanjo, O. A. Breast cancer risk prediction in African women using Random Forest Classifier. *Cancer Treatment and Research Communications*. **2021**, *28*. DOI: 10.1016/j.ctarc.2021.100396
16. Ganggayah, M. D.; Taib, N. A.; Har, Y. C.; Lio, P.; Dhillon, S. K. Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC Medical Informatics and Decision Making*. **2019**, *19* (48). DOI: 10.1186/s12911-019-0801-4
17. Sperandei, S. Understanding logistic regression analysis. *Biochemia Medica (Zagreb)*. **2014**, *21* (1), 12-18. DOI: 1.11613/BM.2014.003
18. Natekin, A.; Knoll, A. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*. **2013**, *7* (21). DOI: 10.3389/fnbot.2013.00021

Author

Rishi Pai is currently a sophomore at Northview High School in Johns Creek, Georgia, USA. He has a deep passion for machine learning, data science, robotics, and materials science/engineering, and hopes to pursue further research projects in these fields.

IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Unraveling Long Non-Coding RNAs in Melanoma: Exploring New Frontiers for Diagnosis and Treatment

Maanya Venkat

Inglemoor High School, 15500 Simonds Rd NE, Kenmore, Washington, 98028, USA; maanyavenkat@gmail.com

ABSTRACT: Melanoma is a brutal skin cancer characterized by its complex properties and high risk of metastasis. Currently, chemotherapy, radiation therapy, and immunotherapy are the most common forms of treatment, however, the persistent rise in melanoma cases and emerging drug resistance underscore the urgency for further effective treatments. Long non-coding RNAs (lncRNAs), once considered 'transcriptional noise', have demonstrated diagnostic and therapeutic biomarker capabilities and salient epigenetic properties to target genes associated with melanoma metastasis and drug resistance. A literature review of 50+ articles from journals and databases, including *Nature*, *Frontiers*, and the *National Library of Medicine*, yielded a combined analysis of the impact of lncRNAs on melanoma diagnosis and therapies. LncRNAs were found to have roles in numerous cellular processes, showcasing their involvement in disease growth and their ability to work, augmenting the effects of current therapies, and ensuring the prevention of melanoma recurrence in the patient. This literature review aims to provide a comprehensive overview of melanoma, examine the functions of lncRNAs in cancers and melanoma, and explore future directions for leveraging lncRNAs to enhance metastatic melanoma treatments and diagnosis.

KEYWORDS: Biomedical and Health Sciences, Genetics and Molecular Biology of Disease, Melanoma Diagnosis and Therapies, Long Non-Coding RNAs.

Introduction

Melanoma, a highly aggressive skin cancer, has seen a staggering 41% increase in melanoma cases globally from 2012 to 2020. The American Cancer Society estimated that in 2021 alone, the US would have about 106,110 new melanoma cases, along with 7,180 deaths, highlighting the ever-urgent need for effective therapies.[1,2] Despite significant advancements in cancer research, current treatments still face the inevitable risk of acquired resistance, resulting in a lack of efficacy. Within the realm of non-coding RNAs, the classification of long non-coding RNAs demonstrates key roles in gene expression regulation, influencing numerous biological processes at a cellular level, including the occurrence and development of diseases such as cancers like melanoma.[3,4] Spanning data from over 50 articles from reputable journals, including *Nature, the International Journal of Molecular Sciences, the National Library of Medicine*, and more, this review aims to explore the current diagnostic and therapeutic approaches for melanoma while investigating how the functions of lncRNAs can improve melanoma treatment and diagnosis strategies.

1. Malignant Melanoma:

Melanoma is a skin cancer resulting from mutations in melanocytes. It mainly affects Caucasians, with common physical characteristics of blue eyes, fair or red hair, pale skin complexion, sunburn history, and freckles.[5,6] In comparison to the Black population, Caucasians have 20 times the risk of melanoma.[2] If not identified in early stages, the melanoma becomes metastatic, resulting in unfortunate prognoses, continued progression of the disease, and increased resistance to therapies.[1,5] To treat cancer, understanding the potential causes of development can help identify drugs to target the appropriate factors. Currently, three main factors act in tandem to determine melanoma development, as further demonstrated by **Table 1**.

Table 1: This table showcases the risk factors and etiology currently known for melanoma. The three main factors include exposure to ultraviolet radiation, atypical mole syndrome, especially with family history, and genetic factors, including gene mutations.

Risk Factor	Description
Lifetime Ultraviolet Radiation	Highly energetic UVB rays (5% of Earth-reaching UV rays) damage DNA in melanocytes in the epidermis, leading to melanoma.[7]
Atypical Mole Syndrome	Also known as dysplastic nevus syndrome, involves abnormal nevi. Carries a 10.7% melanoma risk, increasing to 100% with family history.[6]
Genetics	Family history raises melanoma risk by 2.2 times. 5-10% of patients have a family history. Mutations in genes like CDKN2A, BRAF, and CDK4 are prevalent.[1,2,6,8]

1.1 Melanoma Diagnosis:

To ensure a quick and accurate diagnosis of melanoma, the mnemonic ABCDEF allows both doctors and individuals at home to catch potential cancerous lesions in the early stage, smoothing the diagnosis process exemplified by **Figure 1**. Once a suspicious lesion is identified, an excisional biopsy is typically conducted. After obtaining the results from the biopsy, doctors will look at the blood count, chemistry panel, and lactate dehydrogenase levels to determine the melanoma progression and type.[9]

DOI: 10.36838/v7i8.8

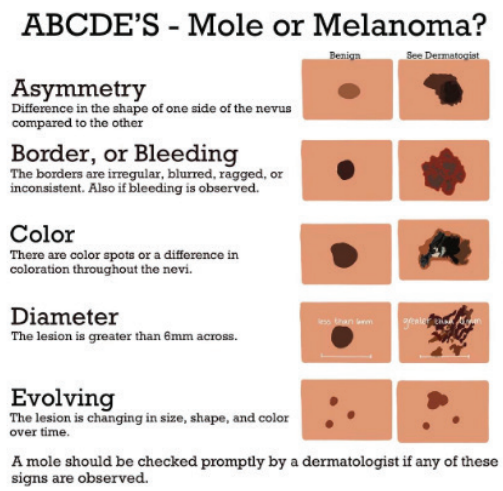


Figure 1: A visual representation of the ABCDE mnemonic that dermatologists use to quickly identify melanoma in a patient. Sometimes, the letter F is also included to represent family history.[9] Copyright 1993-2024, Berman Skin Institute.

1.2 Melanoma Subtypes:

Continuing to build an understanding of melanoma diagnosis, exploring the four main subtypes of melanoma can showcase their differing presentation and progression. **Figures 2-5** showcase the physical properties of each subcategory and provide a detailed description of the statistics and prognosis for each melanoma subtype.

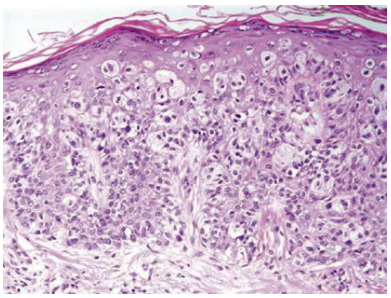


Figure 2: Accounting for over 75% of melanomas, **Superficial Spreading Melanoma** involves unrestricted, singular melanocytes that cause structural modifications in the epidermis with a pagetoid spread.[2,6,10] Superficial Spreading is highly curable, but only when caught early in the progression.[11] Copyright 2006, Bruce R. Smoller, United States & Canadian Academy of Pathology.

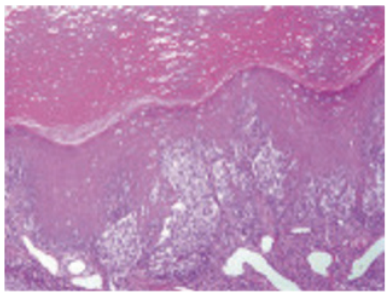


Figure 3: Nodular Melanoma, found in middle-aged adults, has sharply circumscribed tumor edges. Rapid vertical growth and metastasis occur after cancerous melanocytes enter the dermis, resulting in a poor prognosis.[2,6,10] If diagnosed early, it can be cured. However, due to its rapid growth, it is often found at an advanced stage.[12] Copyright 2006, Bruce R. Smoller, United States & Canadian Academy of Pathology.

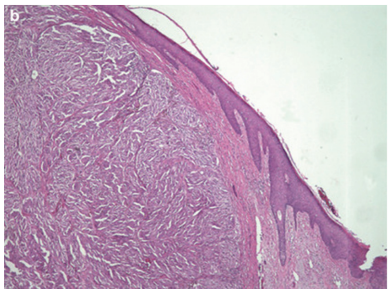


Figure 4: Lentigo Maligna Melanoma, primarily occurring from sun damage, typically forms on the head and neck of elderly people due to small hyperchromatic melanocytes clustering at the dermal-epidermal junction.[2,6,10] If diagnosed early, it can be cured, however, it is often confused with benign sun damage, potentially causing misdiagnosis.[13] Copyright 2006, Bruce R. Smoller, United States & Canadian Academy of Pathology.

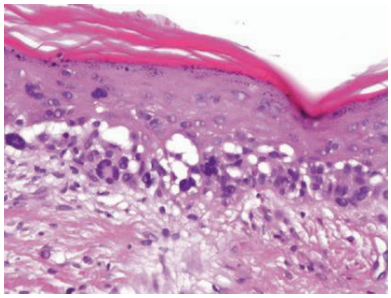


Figure 5: Acral Lentiginous Melanoma is rare and often diagnosed late since it is typically found under the nailbeds It features hyperchromatic melanocytes, both singular and nested, at the dermal-epidermal junction with extensive pagetoid spread.[2,6,10] This is the only melanoma not associated with sun damage and is easily curable if diagnosed early.[14] Copyright 2006, Bruce R. Smoller, United States & Canadian Academy of Pathology.

1.3 Current Melanoma Therapies:

For all subtypes, a high chance of curability is associated with diagnosis and treatment early in the progression of melanoma. Currently, there are three common therapies available to treat melanoma. Before any drugs or treatments are initiated, the cancerous tumor is surgically removed. To increase the overall survival rates, adjuvant therapy is given afterward, such as immunotherapy or targeted therapy. Immunotherapies, such as PD-1 and PD-L1 inhibitor therapies, which suppress T-cell activation, have positive outcomes, with their main role in reducing the occurrences of metastasis.[1,5,9] However, melanoma recurrence is common, as the immune system develops an acquired drug resistance, failing to recognize the T-cells that destroy the cancerous cells. Targeted therapy is a second option as it silences mutated genes in association with melanoma. However, it is still relatively new, and 50% of cases develop acquired resistance to this therapy as well. In cases where the melanoma is too advanced, chemotherapy is the preferred option; Dacarbazine is the standard chemotherapy drug for melanoma. Though it is an important palliative treatment and there are improved clinical responses, there isn't necessarily an improved overall survival for the patient, as the drug could ultimately develop resistance to cellular apoptosis, causing a reappearance of the melanoma.[1,5] In conclusion, a therapy where drug resistance doesn't occur and can effectively kill melanoma is necessary to see improved survival among growing melanoma cases.

2. Long non-coding RNAs:

Long non-coding RNAs (lncRNAs), found in the nucleus and cytoplasm of cells, are a classification of RNA greater than 200 nucleotides long without protein-coding abilities. Initially, lncRNAs were thought to have seemingly little to no biological function, due to a lack of expression and sequence conservation, but research proved that hypothesis wrong.[3,15]

2.1 Understanding and Targeting LncRNAs:

LncRNAs have been found to have important regulatory functions at epigenetic, transcriptional, and post-transcriptional levels due to their complex structure and mechanisms for expression regulation. Hence, they are also closely related to the occurrence, development, and prevention of diseases, including cancer, and play roles in other cellular processes such as the cell cycle, differentiation, and metabolism.[15,16]

2.1.1 Modes of Action:

LncRNAs have four main molecular modes, or functions, of action. In *signal* mode, the lncRNAs function as enhancers in gene imprinting, thereby changing the chromatin architecture, attracting transcriptional proteins to promote target gene transcription, and influencing signaling pathways under specific conditions. In contrast, in *decoy* mode, lncRNAs act as a decoy to block molecular pathways and suppress pre-apoptotic genes. They bind to proteins with transcriptional regulatory functions and control the activity of molecules and signaling pathways to regulate the activation and inhibition of transcription-related genes. *Guide* mode enables lncRNAs to interact and bind with chromatin-modifying enzymes to direct them to a specific local gene for its activation or repression. Lastly, *scaffold* mode, similar to guide, interacts with RNA-binding factors and proteins to form RNA-protein complexes, which either promote or suppress transcription by recruiting proteins to target the promoter region or by binding to existing gene suppressors. **Figure 6** showcases a visual representation of the above information.[3,17-21] An additional mode, *sponging*, allows lncRNAs to act as "molecular sponges" by binding to microRNAs (miRNAs), preventing their interactions with target mRNAs.[22] **Figure 7** showcases this in a visual representation.

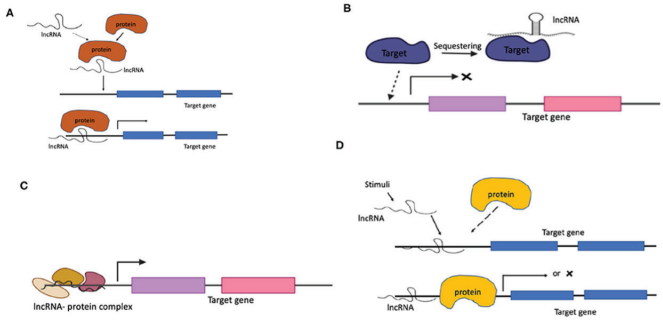


Figure 6: Main modes of action present in lncRNAs. (A) In *signal* mode, the lncRNA interacts with a protein to relay cellular signals to the target gene; (B) The lncRNA sequesters the target enzyme, preventing it from interacting with the target gene in *guide* mode; (C) In *decoy* mode the lncRNA-protein affects the transcription of the target gene; (D) The lncRNA acts as a stimulus for the protein to affect transcription of the target gene in *scaffold* mode.[18] Copyright 2021, Chowdhary *et al.* Luxembourg Centre for Systems Biomedicine.

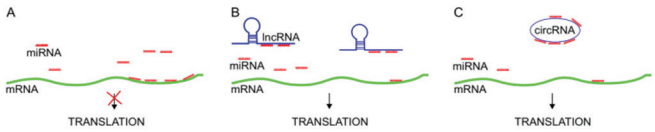


Figure 7: LncRNA sponging pathway. (A) miRNAs bind to the region of target transcriptions to block translation; (B/C) lncRNAs or circRNAs overpower the miRNA, resulting in the resumed translation.[22] Copyright 2021, Wozniak and Czyz, Department of Molecular Biology of Cancer, Medical University of Lodz.

2.1.2 LncRNA-Targeting Biotechnologies:

LncRNAs' modes of action can be targeted to downregulate or upregulate depending on their specific functions. Technologies including antisense oligonucleotides (ASOs), RNA interference (RNAi), and clustered regularly interspaced short palindromic repeats (CRISPR) can be used to aid in the process. ASOs are single-stranded nucleic acids that specifically target lncRNAs and siRNAs for silencing. RNAi uses the cells' own mechanisms for downregulation, and CRISPR can remove lncRNAs entirely.[23] ASOs, specifically, target lncRNAs by degrading lncRNA transcripts. One study using nude mice showed that *in vivo* injected ASOs can inhibit lncRNAs, in this case MALAT1, to block metastasis.[24] With the current knowledge of lncRNA modes of action and overall epigenetic functions, these technologies could be significant in the pursuit of effective cancer therapies utilizing lncRNAs.

2.2 LncRNAs in Cancers:

While continuous research for cancer has led to numerous therapies and diagnosis processes, cancer cases continue to rise on the path to surpassing heart diseases, according to the 2016 American Cancer Statistics Report.[25]

2.2.1 Biomarker Potential:

The use of tumor markers for early diagnosis was identified in 1978. As of right now, tumor antigens, specifically carcinoembryonic antigens (CEAs), glycoproteins, and ectopic hormones, are the most commonly used clinical tumor markers.[25,26] However, due to their extensive applications as a marker, they have a risk of misdiagnosis or the inadvertent neglect of certain markers. Recently, lncRNAs have been discovered as biomarkers for early cancer diagnosis, showing promising results.[25]

LncRNAs such as TINCR (Terminal differentiation-induced non-coding RNA), BANCR (BRAF-activated non-protein coding RNA), and CCAT2 (Colon cancer-associated transcript 2) can be detected from a patient's plasma and function as biomarkers to identify gastric cancer.[18] Exosomal lncRNAs and circulating lncRNAs are promising lncRNAs for oncologists. In addition to their importance in the cell cycle, regeneration, etc., they are involved in tumor growth, metastasis, angiogenesis, and chemoresistance. They are present in bodily fluids, allowing for a plentiful supply.[25,26] Circulating lncRNAs (circRNAs) regulate melanoma growth, invasion, and immune evasion, can stably exist in the circulatory system, and can also be found in bodily fluids. In a study of 32 blood samples from both gastric cancer and healthy patients, the circRNA H19 was expressed more in cancer patients, with-

DOI: 10.36838/v7i8.8

out any evidence of age or race affecting the result; circRNA expression can act as a biomarker. Additional studies showed that the circulating lncRNA MALAT-1 (metastasis-associated lung adenocarcinoma transcript 1) was upregulated in numerous cancer tissues, including but not limited to lung and prostate cancer. MALAT-1 can determine the presence of lung cancer with 96% specificity and prostate cancer with 84.8% specificity.[26] HOTAIR is another lncRNA involved in many carcinogenic processes and is related to the resistance of cisplatin, an antineoplastic agent. It can therefore be a potential biomarker in different cancer cells, such as breast, gastric, colorectal, and cervical cancer cells.[27] Circulating lncRNAs, including H19, HOTAIR, and GACAT2 (gastric cancer-associated transcript 2), have been shown to outperform glycoprotein biomarkers with greater cancer diagnostic performance.[26] Overall, due to their immense involvement in cellular processes and carcinogenesis, lncRNAs have great potential to be used as more widespread biomarkers for diagnosis.

The diverse functions of lncRNAs showcase their incredible potential to be important markers or supplementary treatments in the future. Their established biomarker potential in other cancer types, including gastric cancer, lung cancer, and more, strengthens the thesis that lncRNAs can aid in the early diagnosis of melanoma as well as better curability probabilities. Additionally, with the help of biotechnologies created for gene modification, such as CRISPR and RNAi, lncRNAs could be significant in the fight against drug resistance with current therapies.

3. LncRNAs in Melanoma:

Out of a study using a collection of over 7000 RNA sequencing libraries, 339 lncRNAs were associated with melanoma.[28] LncRNAs can be targeted differently depending on their subtype and functions, be it oncogenic or tumor suppressing, to hold potential for new diagnosis and treatments.

3.1 Functionally Relevant lncRNAs in Melanoma:

A high frequency of lncRNAs prevalent in melanoma are overexpressed, boosting cell proliferation and tumorigenesis.[32] These lncRNAs are oncogenic, as they promote cell growth through the activation of cellular pathways involved in processes like angiogenesis, genomic instability, invasive metastasis, and chemotherapy resistance.[33] Some of the most common oncogenic lncRNAs include H19, SAMMSON, BANCR, HOTAIR, and MALAT1, all of which are upregulated. H19 specifically was the first lncRNA found to be associated with tumorigenesis.[34] Many of these also have sponging abilities and are able to indirectly or directly interact with miRNAs, acting as competing endogenous RNAs (ceRNAs), a specific classification of lncRNAs.[35] Given that oncogenic lncRNAs are a significant group of cellular process regulators and are expressed differently in malignant melanocytes, they have the potential to be biomarkers and used as adjuvant gene therapy to boost the efficacy of current treatments.

A second classification of lncRNAs, tumor suppressors, is downregulated in melanoma and contributes to tumor onset and progression. These lncRNAs can impact regulators in the cell cycle, specifically regulating the expression of tumor suppressor genes to affect proliferation, genomic stability, and apoptosis.[36] Common melanoma-specific tumor suppressor lncRNAs include MEG3 and LINC00961. Many of them specialize in sponging, emphasizing their tumor suppression roles.[22, 28-30] Their abilities to affect cell proliferation, metastasis, growth cycle, and apoptosis highlight promising utilization, especially as a biomarker for melanoma.[29] For further details on the functions and the regulation of specific oncogenic and tumor suppressor lncRNAs, see **Table 2**.

Table 2: This table highlights the main functions and the impact of the regulation of melanoma-specific oncogenic and tumor suppressor lncRNAs. Note: This is not a comprehensive list; only 14 lncRNAs out of many lncRNAs were picked. Additionally, it is important to note that lncRNAs that express both oncogenic and tumor suppressor functions are only found in certain conditions.[33] That said, most of those express oncogenic functions, so that was showcased in the table. * = not explicitly stated in article(s); inferred based on descriptions of lncRNA.

lncRNA	Oncogenic/Tumor Suppressor	Mode of Action	Function	Regulation in Melanoma
BANCR	Oncogenic/Both (ceRNA)[28,32,33]	Sponging[32,37]	BRAF-activated nonprotein coding RNA[28]	Silencing downregulates the MAPK signaling pathway restricting tumor growth and metastasis[28]
CASC2	Tumor Suppressor (ceRNA)[28,32]	Sponging[32]	Promotes tumor growth inhibitor RUNX[22, 28-29]	Downregulated in melanoma; re-expression inhibits melanoma proliferation[22, 28,29]
H19	Oncogenic/Both (ceRNA)[32,33,38]	Sponging (NF-κB pathway), decoy*, signal* [32,39]	Affects melanoma cell growth, metastasis, invasion, and apoptosis; involved in stemness[29,40]	Downregulation regulates signaling pathways restricting melanoma growth[29]
HOTAIR	Oncogenic (ceRNA)[27,28,35]	Sponging (miR-152-3p)[41]	HOTAIR and miR-152-3p sponging suppresses target gene c-MET and activates a signaling pathway to promote melanoma metastasis and growth; involved in chemoresistance[40]	Upregulation of miR-152-3p after the silencing of HOTAIR allowed for lowered cell growth[41]
ILF3-AS1	Oncogenic (ceRNA)[32,38]	Guide* (EZH2 interactions)[42]	Involved in the proliferation and invasion of melanoma cells[52]	Upregulated in melanoma; High levels of ILF3-AS1 corelated with metastasis and poor prognosis. Silencing inhibits cell proliferation[42]
LINC00961	Tumor Suppressor (ceRNA)[28,29,43]	Sponging[43]	Downregulated in melanoma; restriction of cell proliferation, encouragement of apoptosis[22,28-30]	Overall survival with high Linc00961 levels was significantly higher than those with low levels[28,29,43]
MALAT1	Oncogenic/Both (ceRNA) [32,33,38]	Sponging* (miR-23a)[32,44]	Promotion of cell proliferation and migration through miR-23a; involved in angiogenesis[40]	Downregulation of MALAT1 inhibited cell proliferation, migration, expression[44]
MEG3	Tumor Suppressor (ceRNA) [28,29,31,32]	Sponging[31,32,45]	Suppression the growth and pro-apoptosis functions in melanoma[31]	Downregulated in melanoma; low levels of MEG3 were positively correlated with poor prognosis[45]
MIR31HG	Oncogenic[46]	Guide[46]	Upregulated in melanoma; lowered expression of p16INK4A tumor suppressor protein, increased melanoma proliferation[46]	Silencing promotes cellular aging/senescence in the melanoma cells[46]
NKILA	Tumor Suppressor (ceRNA)[32,38]	Decoy* (NF-κB pathway)[32,47]	Inhibition of melanoma invasion and metastasis[28-30]	Downregulated in melanoma; expression triggers apoptosis and reduces melanoma invasion[50]
SAMMSON	Oncogenic[3]	Scaffold* (p32) and Decoy* (CARF protein)[30]	Involved in mitochondrial metabolism, protein translation, and signaling pathways such as Wnt and MAPK[8,30]	Overexpressed in melanoma; silencing weakens the mitochondria's membrane, lessening melanoma survival rate[8,30]
SLNCR	Oncogenic[32]	Guide[32,46]	Promotes p21 expression and contributes to oncogenesis[46]	High levels associated with worse overall survival of melanoma; downregulation significantly decrease melanoma invasion[48]
SPRY4-IT1	Oncogenic[32]	Sponging[32,51]	From the intronic region of SPRY4 gene, associated with BRAF[29]	Silencing induces apoptosis in melanoma[29]
UCA1	Oncogenic (ceRNA)[32,38]	Sponging, Decoy* [8,32]	Promotes melanoma proliferation and invasion through miR-28-5p and miR-507 axes; involved in cell migration and invasion[8,40]	Inhibition of UCA1 expression reduces melanoma cell migration and proliferation; miR-28-5p expression significantly reduced[49]

3.2 LncRNA-based Diagnostic and Treatment Applications:

Currently, lncRNA-based therapies aren't approved for clinical use. However, they are predicted to be crucial in addressing drug resistance and improving the overall survival of melanoma patients as diagnostic and therapeutic targets.

3.2.1 Diagnostic Potential:

LncRNAs have gained attention as diagnostic biomarkers in melanoma for their roles in the regulation and expression of biological processes. Currently, biopsies and physical examinations are the most reliable diagnostic tests for melanoma; however, differentiating between benign and malignant melanocytes can be difficult. LncRNAs would aid as a supplementary biomarker test to further simplify diagnosis.[53] LncRNAs present in melanoma, such as lncRNAs H19, HOTAIR, MEG3, and TUG1, are a few key targets.[54] Though for lncRNAs to be effective diagnostic markers, they should be easily detectable and stable in plasma or other bodily fluids, enabling noninvasive diagnosis.[4]

3.3 Therapeutic Potential:

Rather than a stand-alone treatment, lncRNAs would potentially be used in tandem with current therapies and technologies to increase the chance of overall patient survival.

3.3.1 Immunotherapy:

Immunotherapy, which currently utilizes immune checkpoint inhibitors to enhance T-cell immune responses, lacks efficacy.[29] One recent immunotherapy is the use of PD-1 (programmed cell death protein 1) to treat tumors and to predict patient survival rates. LncRNAs derived from the PD-L1 (PD-1 ligand) gene sites encourage the transcriptional activity of the c-Myc signaling pathway, promoting the growth of cancerous cells. Silencing these lncRNAs and blocking the immune checkpoints could result in tumor suppression. However, despite the potential of immune therapies like anti-PD-1, over half of patients experience drug resistance.[55]

Mechanisms such as the overexpression of the histone methyltransferase EZH2 in melanoma are linked with poor prognosis and affect resistance to the anti-PD-1 treatment.[55] Interestingly, recent studies suggest that tumor-infiltrating immune-related lncRNAs (Ti-lncRNAs) have a higher efficacy rate than the anti-PD-1 treatments when melanoma patients' Ti-lncRNA levels are low.[29] About 15 lncRNAs, including NARF-AS1 and LINC01126, are able to predict prognosis of those treated with anti PD-1. Conversely, lncRNAs, including HOTTIP, IFITM4P, and LINC01140, upregulate PD-L1 expression in melanoma and increase immune cell immune escape.[29, 53] Overall, integrating lncRNAs alongside immunotherapies could enhance their efficacy and restrict resistance challenges.

3.3.2 Targeted Therapy:

Recent studies have linked lncRNAs with BRAF mutations: a significant therapeutic target, affecting over 60% of melanoma patients.[29] Though many BRAF inhibitors (BRAFi) such as vemurafenib and dabrafenib have been approved and widely accepted clinically, resistance to the BRAFi develops after a median time progression-free survival of 9 months.[57, 58] Researchers hypothesize that the repeated exposure of BRAFi stabilizes the modified epigenome in melanoma cells, leading to resistance, assuming such modifications are reversible.

Using this hypothesis, researchers at the University of Colorado Boulder in the Department of Biochemistry conducted a study by treating thousands of singular melanoma cells utilizing BRAFi dabrafenib and observed their reaction over the first four days. While most cells responded positively to the treatment, within three days, a portion of cells escaped the treatment, returning to a cancerous state after drug withdrawal, resulting in acquired resistance.[55,59]

Among the lncRNAs involved in transcriptional activation, EMICERI has been identified as a key player in BRAFi resistance.[56] EMICERI becomes overexpressed following BRAFi resistance, upregulating the MOB3B (MOB kinase activator 3B) gene and downregulating the LATS1 (large tumor suppressor kinase 1) gene. This activates the Hippo signaling pathway, which is largely involved in tumor growth and metastasis due to its functions in tissue and organ growth.[22,58] Targeting EMICERI for downregulation could reduce the effects of the BRAFi resistance, inhibiting further melanoma proliferation.

Another study, conducted by researchers at the VIB-KU Leuven Center for Cancer Biology observed injections of lncRNA SAMMSON antisense nucleotides in mice, which resulted in significant tumor suppression, both independently and with exposure to dabrafenib. However, when dabrafenib was combined with trametinib, the mice endured increased tumor growth. The study determined that SAMMSON would act as a biomarker and a highly selective therapeutic target.[22,60]

In conclusion, lncRNAs such as SAMMSON and EMICERI have proven to significantly aid in overcoming BRAFi resistance, offering promising futures for increased overall patient survival.

Though lncRNAs have not yet been applied to clinical settings, their numerous functions in oncogenic, tumor suppressor, and epigenetic processes highlight their significant potential as biomarkers and adjunctive therapies to address drug resistance and enhance drug efficacy in metastatic melanoma. The extensive studies and research into their prospective implementation within immunotherapies and targeted therapies underscores the promise of lncRNAs in advancing more effective treatments for patients globally.

■ Conclusion

Through this paper, the critical role of lncRNAs in melanoma diagnosis and treatments, focusing primarily on metastasis and drug resistance, has been thoroughly examined. With versatile functions, ranging from diagnostic biomarkers to influencing key cellular processes, including the cell cycle and apoptosis, lncRNAs have proven to present a fascinating and promising pathway for the future of melanoma treatment and diagnosis. Not only can they improve treatment efficacy, but they can also prevent the recurrence of melanoma. Leveraging gene editing biotechnologies, such as CRISPR, can facilitate

DOI: 10.36838/v7i8.8

lncRNAs as an adjuvant targeted therapy to work alongside immunotherapies or chemotherapies. Melanoma continues to remain a formidable territory in oncology, however, continued research on lncRNAs in melanoma is paving the way for clinical implementation. As this research evolves, these molecules, once thought of as mere 'transcriptional noise', hold the potential to revolutionize melanoma care, providing innovative solutions towards the realms of personalized medication, increased patient survival, and perhaps bringing us closer to conquering this deadly cancer.

Acknowledgments

A huge thank you to the following individuals/groups who I could not have completed this project without: First, my mentor, Dr. Hakan Coskun from Harvard Medical School, whose incredible support and belief in my topic fostered an invaluable research experience. Second, the Cambridge Center for International Research staff for providing me with the necessary resources and opportunity to research. Lastly, Prof. Srinivasa N. Raja from Johns Hopkins University for the wonderful guidance and feedback during my final stages of edits.

References

1. Dhanyamraju, P. K.; Patel, T. N. Melanoma therapeutics: a literature review. *Journal of Biomedical Research* **2022**, *36* (2), 77. https://doi.org/10.7555/jbr.36.20210163.
2. Waseh, S.; Lee, J. B. Advances in melanoma: epidemiology, diagnosis, and prognosis. *Frontiers in Medicine* **2023**, *10*. https://doi.org/10.3389/fmed.2023.1268479.
3. Gong, Y.; Zhu, W.; Sun, M.; Shi, L. Bioinformatics Analysis of long non-coding RNA and related diseases: An Overview. *Frontiers in Genetics* **2021**, *12*.
4. Huarte, M. The emerging role of LNCRNAs in cancer. *Nature Medicine* **2015**, *21–21* (11), 1253–1261.
5. Domingues, B.; Lopes, J.; Soares, P.; Populo, H. Melanoma treatment in review. *ImmunoTargets and Therapy* **2018**, *Volume 7*, 35–49. https://doi.org/10.2147/itt.s134842.
6. Heistein, J. B.; Acharya, U.; Mukkamalla, S. K. R. *Malignant melanoma*. StatPearls - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK470409/.
7. Alexander, H. *What's the difference between UVA and UVB rays?* MD Anderson Cancer Center. https://www.mdanderson.org/publications/focused-on-health/what-s-the-difference-between-uva-and-uvb-rays-.h15-1592991.html#:~:text=%E2%80%9CUVB%20does%20not%20penetrate%20as,other%20types%20of%20skin%20cancer.
8. Melixetian, M.; Pelicci, P. G.; Lanfrancone, L. Regulation of LNCRNAs in melanoma and their functional roles in the metastatic process. *Cells* **2022**, *11* (3), 577. https://doi.org/10.3390/cells11030577.
9. *Melanoma treatment*. Cancer.gov. https://www.cancer.gov/types/skin/patient/melanoma-treatment-pdq.
10. Smoller, B. R. Histologic criteria for diagnosing primary cutaneous malignant melanoma. *Modern Pathology* **2006**, *19*, 34–40. https://doi.org/10.1038/modpathol.3800508.
11. Superficial spreading melanoma. Memorial Sloan Kettering Cancer Center. https://www.mskcc.org/cancer-care/types/melanoma/types-melanoma/superficial-spreading-melanoma.
12. *Nodular melanoma.* Memorial Sloan Kettering Cancer Center. https://www.mskcc.org/cancer-care/types/melanoma/types-melanoma/nodular-melanoma.
13. Lentigo maligna melanoma. (n.d.). Memorial Sloan Kettering Cancer Center. https://www.mskcc.org/cancer-care/types/melanoma/types-melanoma/lentigo-maligna-melanoma
14. Acral lentiginous melanoma. (n.d.). Memorial Sloan Kettering Cancer Center. https://www.mskcc.org/cancer-care/types/melanoma/types-melanoma/acral-lentiginous-melanoma.
15. Zhang, X.; Wang, W.; Zhu, W.; Dong, J.; Cheng, Y.; Yin, Z.; Shen, F. Mechanisms and functions of Long Non-Coding RNAs at multiple regulatory levels. *International Journal of Molecular Sciences* **2019**, *20* (22), 5573. https://doi.org/10.3390/ijms20225573.
16. Bridges, M. C.; Daulagala, A. C.; Kourtidis, A. LNCcation: lncRNA localization and function. *The Journal of Cell Biology* **2020**, *220* (2). https://doi.org/10.1083/jcb.202009045.
17. Bhat, S. A.; Ahmad, S. M.; Mumtaz, P. T.; Malik, A. A.; Dar, M. A.; Urwat, U.; Shah, R. A.; Ganai, N. A. Long non-coding RNAs: Mechanism of action and functional utility. *Non-coding RNA Research* **2016**, *1* (1), 43–50. https://doi.org/10.1016/j.ncrna.2016.11.002.
18. Chowdhary, A.; Satagopam, V.; Schneider, R. Long non-coding RNAs: mechanisms, experimental, and computational approaches in identification, characterization, and their biomarker potential in cancer. *Frontiers in Genetics* **2021**, *12*. https://doi.org/10.3389/fgene.2021.649619.
19. Connerty, P.; Lock, R. B.; De Bock, C. E. Long non-coding RNAs: major regulators of cell stress in cancer. *Frontiers in Oncology* **2020**, *10*. https://doi.org/10.3389/fonc.2020.00285.
20. Gao, N.; Li, Y.; Li, J.; Gao, Z.; Yang, Z.; Li, Y.; Liu, H.; Fan, T. Long Non-Coding RNAs: the regulatory mechanisms, research strategies, and future directions in cancers. *Frontiers in Oncology* **2020**, *10*. https://doi.org/10.3389/fonc.2020.598817.
21. Huang, S., Peng, X., Jiang, L., Hu, C. Y., & Ye, W. (2021). LNCRNAs as therapeutic targets and potential biomarkers for Lipid-Related diseases. *Frontiers in Pharmacology*, *12*. https://doi.org/10.3389/fphar.2021.729745
22. Wozniak, M.; Czyz, M. The functional role of Long Non-Coding RNAs in melanoma. *Cancers* **2021**, *13* (19), 4848. https://doi.org/10.3390/cancers13194848.
23. Peng, Q.; Wang, J. Non-coding RNAs in melanoma: biological functions and potential clinical applications. *Molecular Therapy Oncolytics* **2021**, *22*, 219–231. https://doi.org/10.1016/j.omto.2021.05.012.
24. Li, C. H.; Chen, Y. Targeting long non-coding RNAs in cancers: Progress and prospects. *The International Journal of Biochemistry & Cell Biology* **2013**, *45* (8), 1895–1910. https://doi.org/10.1016/j.biocel.2013.05.030.
25. Beylerli, O.; Gareev, I.; Sufianov, A.; Ilyasova, T.; Guang, Y. Long noncoding RNAs as promising biomarkers in cancer. *Non-coding RNA Research* **2022**, *7* (2), 66–70. https://doi.org/10.1016/j.ncrna.2022.02.004.
26. Badowski, C.; He, B.; Garmire, L. X. Blood-derived lncRNAs as biomarkers for cancer diagnosis: the Good, the Bad and the Beauty. *Npj Precision Oncology* **2022**, *6* (1). https://doi.org/10.1038/s41698-022-00283-7.
27. Hajjari, M.; Salavaty, A. HOTAIR: an oncogenic long non-coding RNA in different cancers. DOAJ (DOAJ: Directory of Open Access Journals) **2015**, *12* (1), 1–9. https://doi.org/10.7497/j.issn.2095-3941.2015.0006.
28. De Falco, V.; Napolitano, S.; Esposito, D.; Guerrera, L. P.; Ciardiello, D.; Formisano, L.; Troiani, T. Comprehensive Review on the clinical relevance of Long Non-Coding RNAs in cutaneous mela-

noma. *International Journal of Molecular Sciences* **2021**, *22* (3), 1166. https://doi.org/10.3390/ijms22031166.
29. Zhou, W.; Xu, X.; Cen, Y.; Chen, J. The role of lncRNAs in the tumor microenvironment and immunotherapy of melanoma. *Frontiers in Immunology* **2022**, *13*. https://doi.org/10.3389/fimmu.2022.1085766.
30. Ghasemian, M.; Babaahmadi-Rezaei, H.; Khedri, A.; Selvaraj, C. The oncogenic role of SAMMSON lncRNA in tumorigenesis: A comprehensive review with especial focus on melanoma. *Journal of Cellular and Molecular Medicine* **2023**, *27* (24), 3966–3973. https://doi.org/10.1111/jcmm.17978.
31. Xu, J.; Wang, X.; Zhu, C.; Wang, K. A review of current evidence about lncRNA MEG3: A tumor suppressor in multiple cancers. *Frontiers in Cell and Developmental Biology* **2022**, *10*. https://doi.org/10.3389/fcell.2022.997633.
32. Xiao, Y.; Xia, Y.; Wang, Y.; Xue, C. Pathogenic roles of long noncoding RNAs in melanoma: Implications in diagnosis and therapies. *Genes & Diseases* **2021**, *10* (1), 113–125. https://doi.org/10.1016/j.gendis.2021.08.007.
33. Revathy Nadhan; Dhanasekaran, D. N. Decoding the Oncogenic Signals from the Long Non-Coding RNAs. Onco **2021**, *1* (2), 176–206. https://doi.org/10.3390/onco1020014.
34. Huang, Y.; Linsg, A.; Pareek, S.; Huang, R. S. Oncogene or Tumor Suppressor? Long Noncoding RNAs Role in Patient's Prognosis Varies Depending on Disease Type. Translational Research **2021**, 230, 98–110. https://doi.org/10.1016/j.trsl.2020.10.011.
35. Ma, L.; Bajic, V. B.; Zhang, Z. On the Classification of Long Non-Coding RNAs. RNA biology **2013**, *10* (6), 925–933. https://doi.org/10.4161/rna.24604.
36. Arun, G.; Diermeier, S. D.; Spector, D. L. Therapeutic Targeting of Long Non-Coding RNAs in Cancer. Trends in Molecular Medicine **2018**, *24* (3), 257–277. https://doi.org/10.1016/j.molmed.2018.01.001.
37. Cai, B.; Zheng, Y.; Ma, S.; Xing, Q.; Wang, X.; Yang, B.; Yin, G.; Guan, F. BANCR Contributes to the Growth and Invasion of Melanoma by Functioning as a Competing Endogenous RNA to Upregulate Notch2 Expression by Sponging MiR-204. International Journal of Oncology **2017**, *51* (6), 1941–1951. https://doi.org/10.3892/ijo.2017.4173.
38. LnCeCell 2.0 | Home. Hrbmu.edu.cn. http://bio-bigdata.hrbmu.edu.cn/LnCeCell/LnCeCell2_index.jsp (accessed 2025-05-03).
39. Yang, J.; Qi, M.; Fei, X.; Wang, X.; Wang, K. LncRNA H19: A Novel Oncogene in Multiple Cancers. International Journal of Biological Sciences **2021**, *17* (12), 3188–3208. https://doi.org/10.7150/ijbs.62573.
40. Zhou, W.; Xu, X.; Cen, Y.; Chen, J. The Role of LncRNAs in the Tumor Microenvironment and Immunotherapy of Melanoma. Frontiers in Immunology **2022**, *13*. https://doi.org/10.3389/fimmu.2022.1085766.
41. Luan, W.; Li, R.; Liu, L.; Ni, X.; Shi, Y.; Xia, Y.; Wang, J.; Lu, F.; Xu, B. Long Non-Coding RNA HOTAIR Acts as a Competing Endogenous RNA to Promote Malignant Melanoma Progression by Sponging MiR-152-3p. Oncotarget **2017**, *8* (49), 85401–85414. https://doi.org/10.18632/oncotarget.19910.
42. Chen, X.; Liu, S.; Zhao, X.; Ma, X.; Gao, G.; Yu, L.; Yan, D.; Dong, H.; Sun, W. Long Noncoding RNA ILF3-AS1 Promotes Cell Proliferation, Migration, and Invasion via Negatively Regulating MiR-200b/A/429 in Melanoma. Bioscience Reports **2017**, *37* (6). https://doi.org/10.1042/bsr20171031.
43. Mu, X.; Mou, K.; Ge, R.; Han, D.; Zhou, Y.; Wang, L. Linc00961 Inhibits the Proliferation and Invasion of Skin Melanoma by Targeting the MiR-367/PTEN Axis. International Journal of Oncology **2019**, *55* (3). https://doi.org/10.3892/ijo.2019.4848.
44. Biz, T. B. C.; de Sousa, C.-S. C.; John, S. F.; Galvonas, J. M. LncRNAs in Melanoma Phenotypic Plasticity: Emerging Targets for Promising Therapies. RNA Biology **2024**, *21* (1), 81–93. https://doi.org/10.1080/15476286.2024.2421672.
45. Long, J.; Pi, X. LncRNA-MEG3 Suppresses the Proliferation and Invasion of Melanoma by Regulating CYLD Expression Mediated by Sponging MiR-499-5p. BioMed Research International **2018**, *2018*, 1–15. https://doi.org/10.1155/2018/2086564.
46. Moras, B.; Sissi, C. Unravelling the Regulatory Roles of LncRNAs in Melanoma: From Mechanistic Insights to Target Selection. International Journal of Molecular Sciences **2025**, *26* (5), 2126–2126. https://doi.org/10.3390/ijms26052126.
47. Ouyang, J.; Hu, J.; Chen, J.-L. LncRNAs Regulate the Innate Immune Response to Viral Infection. Wiley Interdisciplinary Reviews: RNA **2015**, *7* (1), 129–143. https://doi.org/10.1002/wrna.1321.
48. Schmidt, K.; Joyce, C. E.; Buquicchio, F.; Brown, A. S.; Ritz, J.; Distel, R.; Yoon, C. H.; Novina, C. D. The LncRNA SLNCR1 Mediates Melanoma Invasion through a Conserved SRA1-like Region. Cell Reports **2016**, *15* (9), 2025–2037. https://doi.org/10.1016/j.celrep.2016.04.018.
49. Han, C.; Tang, F.; Chen, J.; Xu, D.; Li, X.; Xu, Y.; Wang, S.; Zhou, J. Knockdown of LncRNA-UCA1 Inhibits the Proliferation and Migration of Melanoma Cells through Modulating the MiR-28-5p/HOXB3 Axis. Experimental and Therapeutic Medicine **2019**, *7* (5). https://doi.org/10.3892/etm.2019.7421.
50. Wang, Z.; Xie, C.; Chen, X. Diagnostic and Therapeutic Role of Non-Coding RNAs Regulating Programmed Cell Death in Melanoma. Frontiers in Oncology **2024**, *14*. https://doi.org/10.3389/fonc.2024.1476684.
51. Ghafouri-Fard, S.; Khoshbakht, T.; Taheri, M.; Shojaei, S. A Review on the Role of SPRY4-IT1 in the Carcinogenesis. Frontiers in Oncology **2022**, *11*. https://doi.org/10.3389/fonc.2021.779483.
52. Guo, J.-H.; Yin, S.-S.; Liu, H.; Liu, F.; Gao, F.-H. Tumor Microenvironment Immune-Related LncRNA Signature for Patients with Melanoma. Annals of Translational Medicine **2021**, *9* (10), 857–857. https://doi.org/10.21037/atm-21-1794.
53. Masrour, M.; Khanmohammadi, S.; Fallahtafti, P.; Hashemi, S. M.; Rezaei, N. Long non-coding RNA as a potential diagnostic and prognostic biomarker in melanoma: A systematic review and meta-analysis. *Journal of Cellular and Molecular Medicine* **2024**, *28* (3). https://doi.org/10.1111/jcmm.18109.
54. Galus, Ł.; Kolenda, T.; Michalak, M.; Mackiewicz, J. Diagnostic and prognostic role of long non-coding RNAs (lncRNAs) in metastatic melanoma patients with BRAF gene mutation receiving BRAF and MEK inhibitors. *Heliyon* **2024**, *10* (7), e29071. https://doi.org/10.1016/j.heliyon.2024.e29071.
55. Giunta, E. F.; Arrichiello, G.; Curvietto, M.; Pappalardo, A.; Bosso, D.; Rosanova, M.; Diana, A.; Giordano, P.; Petrillo, A.; Federico, P.; Fabozzi, T.; Parola, S.; Riccio, V.; Mucci, B.; Vanella, V.; Festino, L.; Daniele, B.; Ascierto, P. A.; Ottaviano, M. Epigenetic regulation in melanoma: Facts and hopes. *Cells* **2021**, *10* (8), 2048. https://doi.org/10.3390/cells10082048.
56. Proietti, I.; Skroza, N.; Bernardini, N.; Tolino, E.; Balduzzi, V.; Marchesiello, A.; Michelini, S.; Volpe, S.; Mambrin, A.; Mangino, G.; Romeo, G.; Maddalena, P.; Rees, C.; Potenza, C. Mechanisms of acquired BRAF inhibitor resistance in melanoma: a systematic review. *Cancers* **2020**, *12* (10), 2801. https://doi.org/10.3390/cancers12102801.
57. Xu, J.; Wang, X.; Zhu, C.; Wang, K. A review of current evidence about lncRNA MEG3: A tumor suppressor in multiple cancers. *Frontiers in Cell and Developmental Biology* **2022**, *10*. https://doi.org/10.3389/fcell.2022.997633.

DOI: 10.36838/v7i8.8

58. Zhong, J.; Yan, W.; Wang, C.; Liu, W.; Lin, X.; Zou, Z.; Sun, W.; Chen, Y. BRAF inhibitor resistance in melanoma: Mechanisms and alternative therapeutic strategies. *Current Treatment Options in Oncology* **2022**, *23* (11), 1503–1521. https://doi.org/10.1007/s11864-022-01006-7.
59. Yang, C.; Tian, C.; Hoffman, T. E.; Jacobsen, N. K.; Spencer, S. L. Melanoma Subpopulations That Rapidly Escape MAPK Pathway Inhibition Incur DNA Damage and Rely on Stress Signalling. *Nature Communications* **2021**, *12* (1). https://doi.org/10.1038/s41467-021-21549-x
60. Leucci, E.; Vendramin, R.; Spinazzi, M.; Laurette, P.; Fiers, M.; Wouters, J.; Radaelli, E.; Eyckerman, S.; Leonelli, C.; Vanderheyden, K.; Rogiers, A.; Hermans, E.; Baatsen, P.; Aerts, S.; Amant, F.; Van Aelst, S.; van den Oord, J.; de Strooper, B.; Davidson, I.; Lafontaine, D. L. J. Melanoma Addiction to the Long Non-Coding RNA SAMMSON. *Nature* **2016**, *531* (7595), 518–522. https://doi.org/10.1038/nature17161.

Author

Maanya Venkat, a junior at Inglemoor High School, was inspired to explore oncology and genetics following the loss of her family's black labrador. She plans to pursue biomedical engineering in college. Outside of academics, Maanya enjoys tennis, music performance, and working at the local veterinary clinic.

IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Bioinformatics Breakthroughs in Thalassemia: Identifying DDX3 and Potential Drug Leads

Samaya Vaidya[1], Aneesha Vora[1], Nirupma Singh[2]

1) Cathedral and John Connon School, Mumbai, Maharashtra, India; adhithi.arun1304@gmail.com
2) Department of Biological Sciences and Engineering, Netaji Subhas University of Technology (Formerly NSIT), Dwarka, Delhi, India; nirupmajadaun21@gmail.com

ABSTRACT: Thalassemia is a genetic blood disorder that reduces hemoglobin production, leading to oxygen deprivation and severe health complications. Despite the high prevalence of this disease, especially in regions like India, research on potential therapeutic targets has largely focused on the HBB gene, limiting the discovery of effective treatment options. The lack of awareness about this disease remains a major impediment to its early diagnosis and prevention. At present, there are few studies done to discover possible proteins for therapeutic drugs to target, but they have not considered a wide range of genes. We hypothesized that the use of network analysis, pathway analysis, and virtual screening could reveal significantly differentially expressed genes (DEGs) and identify novel therapeutic targets for thalassemia treatment. Five differentially expressed genes were found to gain the most significant positions through network analysis: PRKY, EIF1AY, DDX3Y, CDY2B, and BPY2. Pathway analysis results showed that DDX3 was the top gene present in the top significant pathways that were overrepresented. Hence, DDX3 was found to be an emerging therapeutic target. The binding affinity of different ligands with this DDX3 was computed through virtual screening of large libraries of compounds. Neotetrazolium was found to have the highest binding affinity. Thus, this study contributes to the growing research on therapeutic targets and their counterdrugs for thalassemia and forms the foundation for further experimental validation and development of protein-ligand-based treatments.

KEYWORDS: Computational Biology and Bioinformatics, Proteomics, Thalassemia, Protein-protein interactions, Network and Pathway Analysis, Virtual Screening.

Introduction

Thalassemia is an inherited blood disorder that causes the body to have a lower amount of hemoglobin than what is considered normal due to defects in the synthesis of one or more of the hemoglobin chains.[1, 2] Haemoglobin is a protein found in red cells that carries oxygen from the lungs to all other organs in the body.[3] The body needs it to be able to function. Thalassemia is classified as a trait, minor, intermediate, or major, to describe the severity of the condition.[4] Thalassemia is most commonly found to affect individuals originating from the Mediterranean area, the Middle East, Transcaucasia, Central Asia, North Africa, the Indian subcontinent, and Southeast Asia.[5] India bears a huge burden of hemoglobinopathies, the most prevalent one being thalassemia. Approximately 100,000 Indians are affected by this disease.[6] Thalassemia requires lifelong care, and it was established that children affected with thalassemia need regular blood transfusions, at least twice a month, throughout their lives. The lack of awareness about this disease remains a major impediment to its early diagnosis and prevention.[7]

Thalassemia-affected individuals lack healthy red blood cells, causing a range of problems. When left untreated, severe forms of this disease can cause a change in daily activities and often threaten lives.[8] Thalassemia makes the bone marrow expand, which causes the bones to widen. This can result in abnormal bone structure, especially in the face and skull. Bone marrow expansion also makes bones thin and brittle, increasing the chance of broken bones. Additionally, congestive heart failure and abnormal heart rhythms can be associated with severe thalassemia.[2] Thromboembolic events occur when a thrombus forms and subsequently dislodges. In thalassemia, a common issue called splenectomy can further elevate this risk by increasing circulating platelets, leading to a higher incidence of thrombosis in diseased individuals. Thus, there is a dire need for additional research and early preventative measures to prevent thromboembolic events in thalassemia patients.

Some previous research studies have delved into gene mutations in relation to symptoms of beta-thalassemia. The effects of insertion mutations in the HBB exons have been investigated using the *in silico* approach by using the HbVar database to select sequences with uncharacterized insertion mutations and studying their effects on the structure and function of β-globin protein.[9] In one of the studies, PRMT5 was studied and was found to be a promising target for β-thalassemia, followed by molecular docking to identify its inhibitors.[10] However, none of the previous studies have been found to test a range of genes and proteins.

This study hypothesizes that a broader analysis of gene interactions through differential expression analysis, network analysis, and pathway analysis can identify novel therapeutic targets, which can then be validated for druggability using virtual screening. Network and pathway analysis are sets of widely used tools for research in life sciences intended to give meaning to high-throughput biological data. It is important for

DOI: 10.36838/v7i8.9

diseases such as thalassemia, as the tools analyze data obtained from these high-throughput technologies and detect relevant groups of related genes that are altered in the case of samples in comparison to a control, also known as differentially expressed genes.[11]

We hypothesized that a wider and more in-depth analysis of gene interactions would help discover unique targets that can effectively be used to develop therapeutic drugs for thalassemia treatment. The hypothesis is tested through a three-level validation framework, including DEG analysis, pathway enrichment, and molecular docking. This project introduces unconventional aspects for the study of therapeutic targets for thalassemia. In most cases, studies focus particularly on the HBB gene, whereas this study diverges from the traditional focus on one specific gene. Therefore, the implementation of the novel and innovative framework, which integrates gene expression data with molecular docking, offered fresh insights into therapeutic development for thalassemia. From the study, DDX3 was discovered as the most suitable druggable target for thalassemia, and Neotetrazolium was the best inhibitor, with a binding affinity of -10.2. (ChEMBL ID: 1183691). DDX3Y is located on the Y chromosome and therefore specifically expressed in males; however, DDX3X, its homolog on the X chromosome, may have overlapping functions relevant to both sexes in thalassemia. Thus, this study uses an innovative approach, a combination of differential expression analysis, network, and pathway analysis to identify the significant genes and pathways, and molecular docking to find the best inhibitors, broadening the scope of finding therapeutic drug targets with three-level validation.

Methods

This section presents the hypothesis testing framework of this study in the following sections:

Data Collection:

Gene Expression Omnibus (GEO)[12] was used to gather RNAseq data using different keywords related to thalassemia, such as "Thalassemia," "Beta-thalassemia," and "Haemoglobinopathy." The sample with GSE ID: GSE96060 was chosen from the GEO database to perform differential gene expression analysis of thalassemia and non-thalassemic individuals. The RNASeq dataset GSE96060 comprised samples from individuals diagnosed with β-thalassemia major and healthy controls. Differential gene expression analysis was performed comparing 4 thalassemia samples and 4 non-thalassemic controls. The GEO2R tool was used to identify differentially expressed genes (DEGs) after assigning the 'test' and 'control' groups to each category of samples. The top significantly differentially expressed genes were then downloaded. A library of 12206 drug compounds was downloaded from ChEMBL. The ChEMBL database was selected due to its extensive coverage of bioactive drug-like small molecules, and it has been widely cited in various drug discovery studies. Under 'Small Molecules,' Phase 1, Phase 2, Phase 3, Early Phase 1, and Approved compounds were downloaded. Another library of 21931 compounds was downloaded from the Therapeutic Target Database on PubChem. Similarly, PubChem offers a comprehensive library of small molecules and their biological activities, making it a valuable resource for identifying therapeutic compounds. The RCSB PDB database[13] was used to download the 3D structure of the identified therapeutic target protein.

Network Analysis:

The significant DEGs were then put into STRING (Version 12.0), and protein-protein interaction networks were made by including 5, 10, 20, and 50 interactors subsequently, and 4 networks were downloaded. STRING is a database of known and predicted protein-protein interactions. The networks were visualized using Cytoscape Version 3.10.2 [14] and then analyzed using the AnalyzeNetwork app of Cytoscape. Measures like degree, betweenness centrality, and clustering coefficient were considered to identify the hub genes of the network. The top 25 genes from all 4 networks were identified. The top 10 genes common to all 4 networks and those that were significantly differentially expressed were shortlisted.

Pathway Analysis:

The gene list of the most exhaustive and well-connected network (with 50 added interactors) was put into the Reactome Pathway Database Version 89 [15] to identify the significantly over-represented pathways. The significant pathway results without adding any interactors were downloaded. The top genes of the network were searched for their presence in the significant pathways. The overrepresented and compromised pathways in the case of a patient with Thalassemia were identified from this analysis, which showcased the involvement of top network genes.

Virtual Screening:

In the next phase of the study, from the ChEMBL database, 'Small Molecules' (Phase 1, Phase 2, Phase 3, Early Phase 1, Approved compounds) were downloaded. From PubChem, the 'Therapeutic Target Database' library was downloaded. DDX3 was prepared for docking using Autodock.[16] The libraries of ligands were virtually screened against DDX3 as the target protein, using PyRx software. [17] The AutoDock Vina scoring function was used to compute binding affinities. The binding affinity of each ligand was computed, and the top compounds from both libraries were identified. BIOVIA Discovery Studio was used to visualize the interaction maps of the top 3 ligands, each from TTD and ChEMBL libraries, with DDX3.

Result and Discussion

This section presents the results obtained from the bioinformatics analysis of differentially expressed genes, their pathways, and network interactions, providing a wider and more in-depth analysis of gene interactions in Thalassemia and testing the hypothesis. The analysis helped discover unique targets with a focus on understanding the potential drugs associated with therapeutic targets of Thalassemia.

Differential gene expression analysis:

After identifying the RNASeq dataset GSE96060, retrieved from the GEO, 20,047 differentially expressed genes were retained for analysis. A p-value threshold of <0.05 and a fold-change of ±2 were applied in GEO2R. The expression of DEGs was visualized in the form of a volcano plot (Figure 1). In total, 23 genes were found to have significant differential gene expression with a p-value <0.05 and a fold change of ±2 and were considered statistically significant. The list of these genes was used for further analysis.

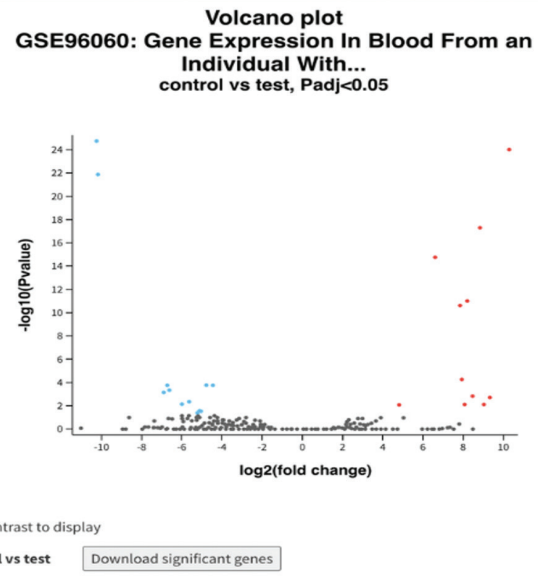


Figure 1: The volcano plot of differentially expressed genes in Thalassemia. This plot shows the gene expression in blood from control and test samples with Thalassemia, identified using the GEO2R tool. This helped in identifying the significant genes that were differentially expressed between normal individuals and thalassemia patients.

Network and Pathway Analysis:

Network study helped in visualizing the interaction maps of significantly differentially expressed genes of all 4 networks for 5, 10, 20, and 50 interactors (Figure 2). The network with 50 added interactors showed 4 clusters of connected components, while the network with 20 added interactors showed the most connected components in one network. The 50 interactors were the most exhaustive network in terms of connectivity and in showing the thalassemia disease network components; thus, they were used to perform pathway analysis. Networks were analyzed based on degree, betweenness centrality, and clustering coefficient. These are the topological measures of a network that help identify the top central genes, which are well-connected and are the hub genes of the network. The genes in all 4 networks that had the highest of all three centrality measures and those that were significantly differentially expressed were PRKY, EIF1AY, DDX3Y, CDY2B, and BPY2. These genes secured the most significant position in the disease network in thalassemia.

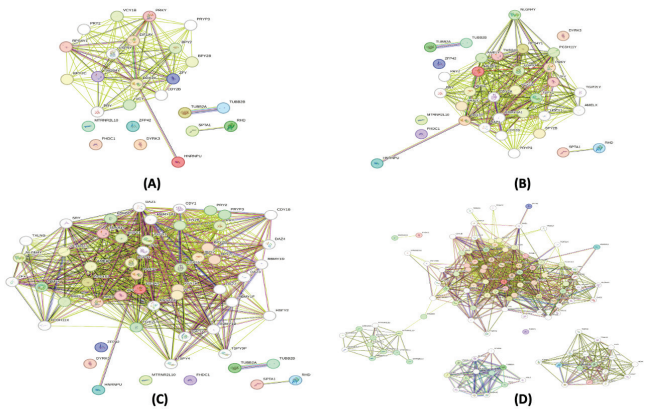


Figure 2: The interaction network of differentially expressed genes. (A) This shows an interaction network of differentially expressed genes with 5 added interactors in STRING. (B) This shows an interaction network of differentially expressed genes with 10 added interactors in STRING. (C) This shows an interaction network of differentially expressed genes with 20 added interactors in STRING. (D) This shows an interaction network of differentially expressed genes with 50 added interactors in STRING. Networks helped in identifying the well-connected hub genes of the thalassemia disease network, which were further cross-checked to be differentially expressed.

Pathway analysis using the Reactome Pathway database showed 47 over-represented pathways in the thalassemia disease network; those were found to be statistically significant with a p-value <0.05. The top enriched pathways in this analysis were key biological processes, such as protein synthesis, RBC membrane integrity, and immune responses, all of which play roles in the pathology of thalassemia. The blood group biosynthesis and translation initiation pathways show the disruptions in hemoglobin production and indicate the need for transfusion in thalassemia patients. These findings suggest a complex interplay between genetic mutations in hemoglobin genes, alterations in protein synthesis pathways, and the body's response to defective RBCs, all contributing to the clinical manifestations of thalassemia. The top genes of the network were searched in the significantly over-represented pathways to check their involvement in the top thalassemia-related pathways. DDX3Y and EIF1AY were found to be present in the top 10 significant pathways, thus indicating their prime role in thalassemia disease.

These two genes were screened through the literature for the existing evidence of their role in Thalassemia and their druggability role. It was identified that both DDX3Y and EIF1AY are suitable as drug targets, but DDX3Y was identified to have a stronger association and prime role in thalassemia than EIF1AY (12). Thus, amongst the top genes, DDX3Y was found to be the most suitable therapeutic target protein as it passed three levels of validation, which are holding the top position in the network, are involved in top significant pathways, and are backed by the literature for the involvement in thalassemia disease complications. DDX3 also showed significant differential gene expression in thalassemia patients as compared to non-thalassaemic individuals (Figure 3). These findings support the hypothesis that DDX3Y is a critical player in thalassemia's molecular landscape. DDX3 was thus

DOI: 10.36838/v7i8.9

selected for further analysis as a potential therapeutic target to identify its inhibitors.

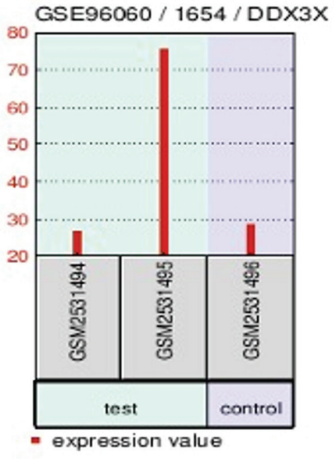


Figure 3: Differential expression of DDX3 in Thalassemia patients. The red bar in the bar graph shows the differential expression (over-expression in homozygous sample) of DDX3 in test samples of patients with Thalassemia as compared to the normal control samples. DDX3 was found to be significantly differentially expressed in thalassemic patients and was also found to be the top hub gene in the disease network.

Virtual Screening analysis:

After performing virtual screening of the libraries of compounds from PubChem and ChEMBL, the top 6 ligands from PubChem (13) TTD and ChEMBL (14) databases were identified based on binding affinity towards DDX3 (Table 1). Neotetrazolium with ChEMBL id CHEMBL1183691 was identified to have the highest binding affinity of -10.2. The lower the binding affinity, the higher the strength of the interaction. Thus, Neotetrazolium holds the most significant position in exhibiting the best binding affinity to DDX3. The surface interaction of Neotetrazolium was visualized in a docked position with DDX3 (Figure 4), along with the interaction of the compound with DDX3 in the ribbon cartoon structure (Figure 5). The amino acid residue level interaction of Neotetrazolium with DDX3 was identified using Biovia Discovery Studio. The interaction of residues of A and B chains of DDX3 with the Neotetrazolium compound was visualized (Figure 6). The compound was seen to form 1 hydrogen bond, 2 pi-cation bonds, 1 pi-pi stacking interaction, 2 pi-alkyl interactions, and 1 pi-sigma interaction with the A and B chains of DDX3.

Table 1: The binding affinities of the top 6 compounds from ChEMBL and PubChem libraries. RMSD/ub and RMSD/lb refer to upper and lower bound root mean square deviation, respectively.

	Compound ID and name	Binding Affinity	rmsd/ub	rmsd/lb
	CHEMBL1183691 (NEOTETRAZOLIUM)	-10.2	0	0
ChEMBL	CHEMBL1079593 (VS-5584)	-10	0	0
	CHEMBL1076263 (SETROBUVIR)	-9.2	0	0
	10393120 (manzamine Y|D05GIC)	-9.9	0	0
PubChem	10029385 (LY2090314|D0Z1DH)	-9.8	0	0
	10218379 (NDT9520492|D0J3FH)	-9.8	0	0

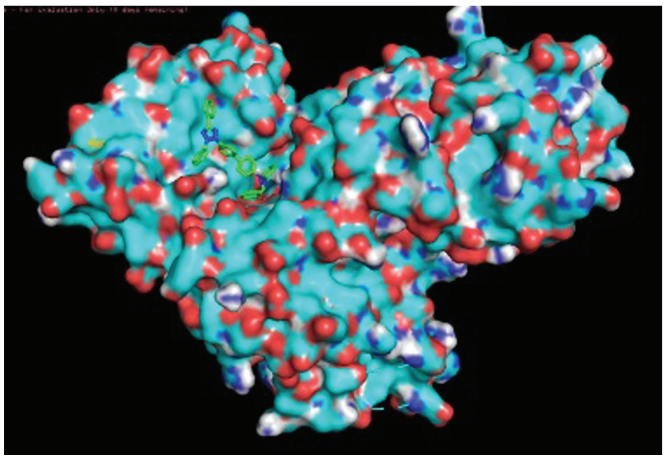


Figure 4: Neotetrazolium docked with DDX3 in surface structure. This shows the top-ranked ligand, Neotetrazolium ligand in green color, docked with DDX3 (surface shown in blue and red color), the most significantly differentially expressed gene in thalassemia. This helped in visualizing the perfect binding of Neotetrazolium in the surface pockets of DDX3.

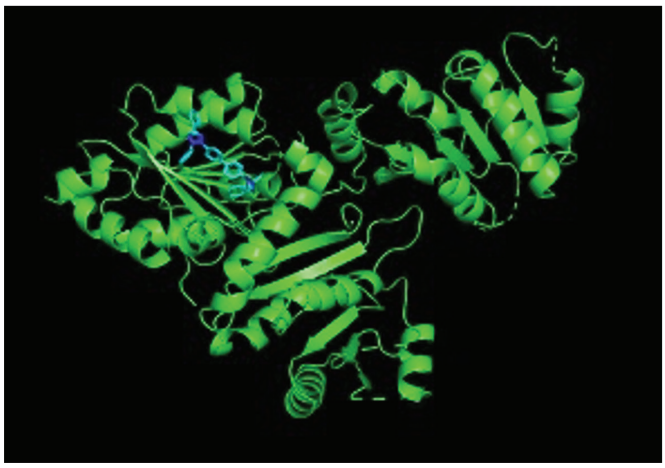


Figure 5: Neotetrazolium docked with DDX3 in ribbon form. This shows the top-ranked ligand, Neotetrazolium ligand in blue color, docked with DDX3 in green color, the most significantly differentially expressed gene.

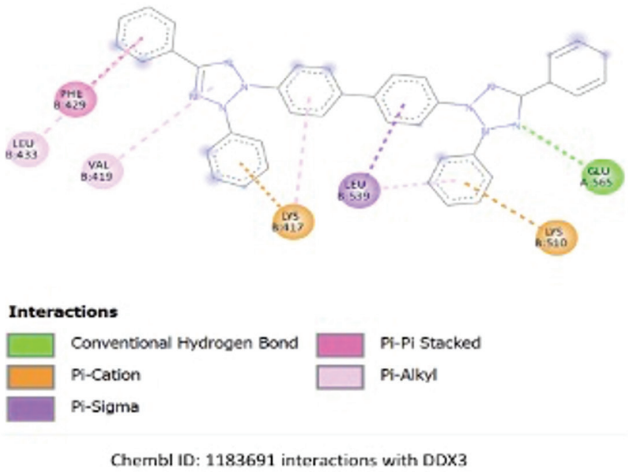


Figure 6: Interactions of Neotetrazolium with DDX3. This shows the amino acid residue interactions of the top-ranked ligand, Neotetrazolium, with DDX3. The hydrogen bond and π-interactions identified contribute to the binding strength and stability of the DDX3–Neotetrazolium complex. The rectangular color bars show the color coding of each type of interaction. This helped in identifying the chemical interactions between the protein and the top-ranked ligand.

Prior research on thalassemia has primarily focused on HBB. Therefore, the goal of this study was to go beyond traditional genes like HBB and discover new therapeutic targets and lead compounds through network and pathway analysis and virtual screening. Thus, it was hypothesized, through our key question, that a wider and more in-depth analysis of gene interactions would help discover unique targets that can effectively be used to develop well-analyzed and researched therapeutic drugs that will uncover new insights into Thalassemia treatment. The results validate the hypothesis that a systems-level approach can identify novel therapeutic targets for thalassemia. DDX3 was discovered as a therapeutic druggable target for thalassemia, and Neotetrazolium is the best ligand that can effectively bind to the target. DDX3 is also a proposed therapeutic target for lung cancer, which indicates its suitability as a drug target.[18]

Through network analysis using STRING,[19] the genes holding the top position in all 4 networks and those that were significantly differentially expressed were PRKY, EIF1AY, DDX3Y, CDY2B, and BPY2. Amongst the top genes, DDX3 was selected for further investigation as it was found to be the most suitable druggable protein in the literature. It was found to be involved in the top over-represented significant thalassemia-related pathways, was present in significantly differentially expressed genes, and held the top position in the network, i.e., it had a high degree, high betweenness centrality, and high clustering coefficient. The results indicated that DDX3 was significantly upregulated in individuals with Thalassemia, as seen in other diseases, such as Medulloblastoma.[20] Gene enrichment analysis revealed that the Rhesus Blood Group Biosynthesis Pathway is heavily involved in the progression of Thalassemia, providing new insights into potential therapeutic targets and negating HBB as the primary gene studied in such cases.[21] Next, the top 6 ligands from PubChem, TTD, and ChEMBL databases were identified based on binding affinity towards DDX3, where Neotetrazolium had the highest binding affinity of -10.2. (ChEMBL ID: 1183691). These findings also highlight the potential for re-purposing existing DDX3-targeted therapies or developing new agents for treating hematologic disorders such as thalassemia. The schematic of the whole workflow is shown in Figure 7.

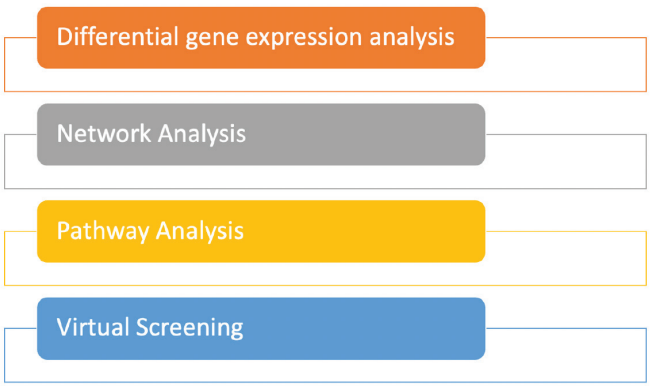


Figure 7: Schematic of the study workflow showing four phases. 1. Differential Gene Expression Analysis, 2. Network Analysis 3. Pathway Analysis and 4. Virtual Screening.

The research conducted, therefore, not only contributes to existing knowledge of potential therapeutic drug targets to inhibit the expression of thalassemia but also adds to it through the novel discovery of the potential use of DDX3. The gene DDX3 was originally identified through this gene expression and possibly would not have been studied if solely focusing on the most well-known genes, such as HBB, that manifest in disease pathways. The identification of DDX3 as a potential diagnostic marker suggests its utility in developing early detection methods for thalassemia, possibly even before symptoms manifest. A major strength of this study is the comprehensive use of multiple bioinformatics tools, including a multitude of datasets from PubChem and ChEMBL, which allowed for an in-depth analysis of the genomic data, leading to the eventual discovery of DDX3 as the most promising target.[22] Targeted therapies against DDX3 could potentially complement existing thalassemia management approaches, including transfusion protocols, iron chelation therapy, and emerging gene therapies. The integration of pathway and network analysis provided a holistic view of the underlying biological mechanisms. Furthermore, virtual screening and molecular docking served as further validation of an already sound workflow. Thus, the study depicts the potential of advanced bioinformatics tools in the identification of novel therapeutic targets for Thalassemia. The interaction between the DDX3 and identified potential ligands, such as Neotetrazolium, can be further experimentally validated for their suitability in the clinical setting. The hypothesis-driven approach used in this study provides a framework for future research, including experimental validation of identified targets and inhibitors.

Conclusion

Overall, the findings of this study contribute to establishing a foundation for further research, as they underscore the sheer importance of bioinformatics in exploring disease mechanisms as well as discovering new drug targets. The importance of this study extends far beyond just Thalassemia. It serves as an example of how these various approaches can uncover targets that have been previously overlooked. As bioinformatics continues to evolve, studies like this demonstrate the potential of combining computational methods with experimental research to speed up medicinal discoveries. With more research and experimental validation, DDX3 could become a fundamental principle in Thalassemia treatment strategies. The application of bioinformatics techniques continues to open new avenues for research, with the potential to improve personalized medicine and therapeutic interventions.

Acknowledgments

SV and AV would like to acknowledge Aashna Saraf, Founder of CreatED, for providing valuable feedback and guidance throughout the project.

References

1. Muncie, H. L., Jr.; Campbell, J., Alpha and beta thalassemia. *Am Fam Physician* **2009**, *80* (4), 339-44.

DOI: 10.36838/v7i8.9

2. Mayo Clinic Thalassemia. https://www.mayoclinic.org/diseases-conditions/thalassemia/symptoms-causes/syc-20354995 (accessed 27 August).
3. Marengo-Rowe, A. J., Structure-function relations of human hemoglobins. *Proc (Bayl Univ Med Cent)* **2006**, *19* (3), 239-45.
4. He, L. N.; Chen, W.; Yang, Y.; Xie, Y. J.; Xiong, Z. Y.; Chen, D. Y.; Lu, D.; Liu, N. Q.; Yang, Y. H.; Sun, X. F., Elevated Prevalence of Abnormal Glucose Metabolism and Other Endocrine Disorders in Patients with β-Thalassemia Major: A Meta-Analysis. *Biomed Res Int* **2019**, *2019*, 6573497.
5. Cao, A.; Kan, Y. W., The prevention of thalassemia. *Cold Spring Harb Perspect Med* **2013**, *3* (2), a011775.
6. Colah, R.; Italia, K.; Gorakshakar, A., Burden of thalassemia in India: The road map for control. *Pediatric Hematology Oncology Journal* **2017**, *2* (4), 79-84.
7. Ali, S.; Mumtaz, S.; Shakir, H. A.; Khan, M.; Tahir, H. M.; Mumtaz, S.; Mughal, T. A.; Hassan, A.; Kazmi, S. A. R.; Sadia; Irfan, M.; Khan, M. A., Current status of beta-thalassemia and its treatment strategies. *Mol Genet Genomic Med* **2021**, *9* (12), e1788.
8. Shi, P. A., Transfusion Management in Patients with Hemoglobinopathies. In *Transfusion Medicine and Hemostasis (Second Edition)*, al, S. e., Ed. Elsevier: San Diego, 2013; pp 327-336.
9. Qadah, T.; Jamal, M. S., Computational Analysis of Protein Structure Changes as a Result of Nondeletion Insertion Mutations in Human β-Globin Gene Suggests Possible Cause of β-Thalassemia. *Biomed Res Int* **2019**, *2019*, 9210841.
10. Pokharel, B.; Ravikumar, Y.; Rathinavel, L.; Chewonarin, T.; Pongpom, M.; Tipsuwan, W.; Koonyosying, P.; Srichairatanakool, S., The Discovery of Selective Protein Arginine Methyltransferase 5 Inhibitors in the Management of β-Thalassemia through Computational Methods. *Molecules* **2024**, *29* (11), 2662.
11. García-Campos, M. A.; Espinal-Enríquez, J.; Hernández-Lemus, E., Pathway Analysis: State of the Art. *Front Physiol* **2015**, *6*, 383.
12. Barrett, T.; Edgar, R., Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol* **2006**, *411*, 352-69.
13. Zardecki, C.; Dutta, S.; Goodsell, D. S.; Voigt, M.; Burley, S. K., RCSB Protein Data Bank: A Resource for Chemical, Biochemical, and Structural Explorations of Large and Small Biomolecules. *Journal of Chemical Education* **2016**, *93* (3), 569-575.
14. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T., Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **2003**, *13* (11), 2498-504.
15. Fabregat, A.; Sidiropoulos, K.; Viteri, G.; Forner, O.; Marin-Garcia, P.; Arnau, V.; D'Eustachio, P.; Stein, L.; Hermjakob, H., Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinformatics* **2017**, *18* (1), 142.
16. Santos-Martins, D.; Forli, S.; Ramos, M. J.; Olson, A. J., AutoDock4(Zn): an improved AutoDock force field for small-molecule docking to zinc metalloproteins. *J Chem Inf Model* **2014**, *54* (8), 2371-9.
17. Dallakyan, S.; Olson, A. J., Small-Molecule Library Screening by Docking with PyRx. In *Chemical Biology: Methods and Protocols*, Hempel, J. E.; Williams, C. H.; Hong, C. C., Eds. Springer New York: New York, NY, 2015; pp 243-250.
18. Bol, G. M.; Vesuna, F.; Xie, M.; Zeng, J.; Aziz, K.; Gandhi, N.; Levine, A.; Irving, A.; Korz, D.; Tantravedi, S.; Heerma van Voss, M. R.; Gabrielson, K.; Bordt, E. A.; Polster, B. M.; Cope, L.; van der Groep, P.; Kondaskar, A.; Rudek, M. A.; Hosmane, R. S.; van der Wall, E.; van Diest, P. J.; Tran, P. T.; Raman, V., Targeting DDX3 with a small molecule inhibitor for lung cancer therapy. *EMBO Molecular Medicine* **2015**, *7* (5), 648-669-669.
19. Mering, C. v.; Huynen, M.; Jaeggi, D.; Schmidt, S.; Bork, P.; Snel, B., STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research* **2003**, *31* (1), 258-261.
20. Swarup, A.; Bolger, T. A., The Role of the RNA Helicase DDX3X in Medulloblastoma Progression. *Biomolecules* **2024**, *14* (7), 803.
21. Thein, S., Molecular basis of β thalassemia and potential therapeutic targets. *Blood Cells, Molecules, and Diseases* **2017**, *70*.
22. Rosenkranz, A. A.; Slastnikova, T. A., Prospects of Using Protein Engineering for Selective Drug Delivery into a Specific Compartment of Target Cells. *Pharmaceutics* **2023**, *15* (3)

Authors

Samaya is an 11th grader studying the International Baccalaureate Diploma Program at The Cathedral and John Connon School, Mumbai. She wrote this paper at the end of the 10th grade, leading into the 11th grade. Now, she works towards giving young teenagers in her school the opportunity to engage in research in STEM to garner a holistic view of what it means to be a person studying the sciences with the launch of her website, 'To The Core.' In the future, she hopes to explore the field of Medicine.

Aneesha is an 11th grader studying the International Baccalaureate Diploma Program at The Cathedral and John Connon School, Mumbai. Being extremely passionate about Mathematics and Biology, she undertook a Program in Bioinformatics in the Summer of 2024. This formed the basis of the research paper demonstrating her curiosity in this area. Going ahead, she is keen to pursue Biology for her undergraduate studies.

Nirupma Singh is a Bioinformatics Scientist with a doctorate in Biotechnology and Bioinformatics from the University of Delhi with six years of hands-on research and development experience. Her journey is marked by a robust foundation in Machine Learning and Python, with five years of expertise. Proficient in Linux and cloud servers like AWS. She excels in structural biology, systems biology, protein/gene network analysis, data mining, and computational genomics.

IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

A Novel Deep Learning Gut Microbial Analysis for Biomarker Identification and Diagnosis of IBS

Anav Gupta[1], Nirupma Singh[2]

1) The Shri Ram School Moulsari, V37 Moulsari Avenue, Gurugram, Haryana, 122002, India.
2) Department of Biological Sciences and Engineering, Netaji Subhas University of Technology (Formerly NSIT), Dwarka, Delhi, India.; nirupmajadaun21@gmail.com

ABSTRACT: Irritable Bowel Syndrome (IBS) is a widespread functional disorder of the gastrointestinal tract (GI), characterized by abdominal pain and altered bowel habits. Traditional IBS diagnosis is done through self-reported questionnaires characterizing frequencies and severities of associated symptoms. The gut microbiome has shown a significant correlation with various IBS symptoms, and treatments targeting it have been effective in mitigating its symptoms. Previous studies have correlated the gut microbiome with a single label of IBS, whereas it has multiple subtypes. In this study, machine learning (ML) and deep learning (DL) models have been developed to correlate gut microbiome with five IBS symptoms: constipation, acidity, diarrhea, bloating, and burping. Using data from 1100 patients, Graph Neural Networks (GNNs) outperformed traditional ML models by ~15%, while Feed Forward Neural Networks (FFNNs) showed ~20% improvement, achieving 85-90% accuracy for symptom severities and 80-85% for frequencies. Permutation importance was used to compute the important features according to the model, along with Pearson's correlation analysis to identify the direction in which the features varied with the output. Taxon 1737404 (*Murdochiella vaginalis*), 768507 (*Runella slithyformis*), and 1760 (*Actinomyces israelii*) had the highest permutation scores with a positive correlation, while 1236 (*Escherichia coli*) and 186826 (*Enterococcus faecium*) had the strongest permutation scores with a negative correlation. Two web applications were developed for the model, one of which allows other clinicians to upload their datasets, and the other, which returns predictions based on uploaded gut tests. Thus, this study demonstrates the potential of deep learning to leverage gut microbiome data for the accurate prediction of IBS symptoms, along with identifying essential biomarkers.

KEYWORDS: Computational Biology and Bioinformatics, Computational Biomodelling, Irritable Bowel Syndrome, Gut Microbiome Biomarker identification and diagnosis.

Introduction

IBS is a collection of long-term digestive conditions that are prevalent in the GI tract.[1] One in every eleven people globally and one in every fourteen people in India suffer from a set of conditions known as IBS.[2] IBS significantly impacts quality of life, causing chronic abdominal pain, irregular bowel movements, and psychological distress that can affect work productivity and social relationships. Existing literature has shown an imbalance in the gut microbiome, or "gut dysbiosis", to have a strong correlation with IBS.[3] Many therapies that tailor to the gut microbiome have shown a significant improvement in overall IBS Symptoms.[4] The current diagnosis of IBS is based on the Rome IV criteria, which involves clinicians studying self-reported severities and frequencies of various IBS-related symptoms from patients and using that data to diagnose them with IBS.[5] Some of these symptoms include bloating, acidity, constipation, diarrhea, and burping. As per the literature review, there are no reliable biomarkers for IBS.

Research has shown an exponential growth in microbial data, outlining the need for ML.[6] ML models have been developed by correlating the gut microbiome with a single IBS output, showing significant correlations.[7] ML and DL have also been used extensively in microbiome research, with studies showing DL to be more powerful due to its ability to capture microbial dynamics.[8] This research study aims to develop a software and a platform to perform IBS diagnosis based on information gained from the gut microbiome, using symptom-based classification of symptoms. Traditional IBS models, although rare, are trained on a binary IBS output column. This study aims to widen the scope of IBS diagnosis by training the model using the various IBS symptoms (their self-reported severities and frequencies). Testing for severities and frequencies could aid the clinical landscape for IBS, as clinicians can identify targeted biomarkers for each symptom, and subjectivity in patient responses can be eliminated.

Previous microbiome-IBS studies have trained models based on Caucasian and American datasets (such as the European Nucleotide Archive and the American Gut Project). [9, 10] Since the gut microbiome is influenced by the environment, a South Asian/ Indian microbial profile is likely very different from an American/ Caucasian profile. Thus, in this study the relative abundance data of the taxon in the gut microbiome for 1089 Indian patients was gathered with a questionnaire containing how badly (on a scale of 1-10), symptoms such as bloating, acidity, constipation, diarrhea, and burping affected their daily lives and how frequently (per week) these symptoms affected them). Several ML models, including Random Forest (RF) and Adaptive Boosting (AB), were run to build the models.

DOI: 10.36838/v7i8.10

A Graph Convolutional Network (GCN) was also developed, followed by FFNNs, which combine the GNN's ability to effectively capture the taxa in the gut with a simpler architecture, delivering a significantly higher accuracy (of around 85-90% for severities and 80-85% for frequencies). Feature importances were also identified. Overall, this work represents a significant step towards objective, microbiome-based IBS diagnosis and targeted treatments while addressing the critical gap in population-specific research for the Indian subcontinent.

Methods

Data Collection:

1089 Gut Microbiome Tests (GMTs) were sourced from SOVA Health. Microbiome data were derived from 16S rRNA sequencing and provided as relative abundance profiles. These GMTs were collected anonymously, with a unique identifier for each patient, adhering to the ethics of data privacy. Each of these GMTs contained the relative abundances of 31258 taxon and a self-reported questionnaire by patients based on the Rome IV criteria of IBS regarding the severities and frequencies of various IBS symptoms. All the data was anonymized and collected in compliance with data privacy regulations.

Model Building:

Ten models of each type were trained for different symptom severities and frequencies. RF and AB ML models were developed and compared. For the deep learning approach, a GNN was built using Torch Geometric. The network architecture consisted of three Graph Convolutional Network (GCN) layers followed by global pooling and fully connected layers. The algorithm constructs a weighted graph representing relationships between microbiome features based on their co-occurrence patterns in the dataset. Initially, it computes individual feature support values by calculating the proportion of samples where each feature is present. The pairwise relationships between features are then quantified using two complementary metrics: lift and Jaccard similarity. The lift metric measures how much more likely features are to occur together compared to random chance, while the Jaccard similarity captures the overlap between feature occurrences relative to their union. These metrics are combined using a weighted average (70% lift, 30% Jaccard) to produce a final co-occurrence score for each feature pair. The algorithm then creates graph edges for feature pairs whose combined score exceeds a specified threshold (default 0.3). To ensure the graph remains fully connected and to stabilize subsequent graph-based computations, self-loops with maximum weight (1.0) were added for each feature. The resulting graph structure was represented using two Torch tensors: an edge index tensor encoding the connectivity pattern and an edge weight tensor containing the corresponding co-occurrence scores, as shown in Figure 1. This graph representation captured both direct and indirect relationships between microbiome features, enabling the model to leverage community structure information during learning.

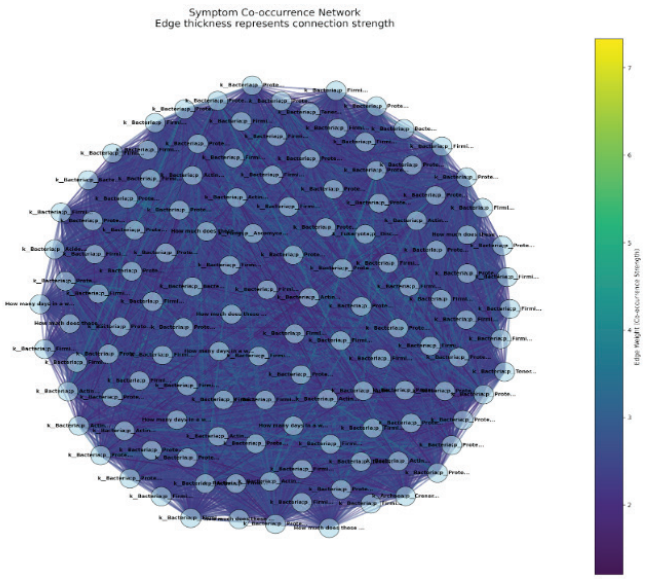


Figure 1: A graph visualisation, created for an individual sample. Round, light blue nodes represent taxon, and edge connections represent co-occurrences between taxon, ranging from violet to light blue: violet having the least weight and yellow having the most weight.

For activation functions, ReLU was used after each GCN layer, followed by batch normalization and a dropout rate of 0.3 to prevent overfitting. The model combined both mean and sum pooling operations to capture different aspects of the graph structure. The final classification layers used ReLU activation with a softmax output layer for multi-class prediction. The GNN implemented a graph convolutional network architecture designed for symptom classification, as shown in Figure 2. It employed three sequential GCN layers (GCNConv) that transformed the input features through message passing operations across the graph structure. Each GCN layer mapped the input to a hidden dimension space, maintaining the same hidden dimensionality across layers. The architecture incorporated batch normalization after each convolution to stabilize training and accelerate convergence. Following the graph convolutions, the model combined global mean and sum pooling operations to aggregate node-level features into graph-level representations. These pooled features were concatenated and processed through two fully connected layers for final classification. Dropout (0.3) is applied throughout to prevent overfitting. The model uses ReLU activation functions between layers to introduce non-linearity. This architecture enables the network to learn both local structural patterns through convolutions and global graph properties through pooling operations.

The following were the hyperparameters that were used for tuning.

Learning rate: [0.01, 0.001, 0.0001]
Hidden Layer Dimensions: [64,128,256]
Dropout rate: [0.2,0.3,0.4]
Co-occurrence threshold: [0.2, 0.3, 0.4]

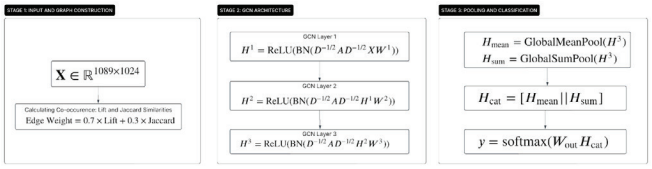


Figure 2: The architecture of the GNN. This figure shows a flowchart of its components, and mathematical functions are utilized to construct the graphs and train the model.

A specialized FFNN architecture, as shown in Figure 3, was developed using PyTorch Geometric, comprising three interconnected components. The proposed FFNN architecture consists of three parallel processing pathways that capture different aspects of the microbiome data. The first pathway implements a feature attention mechanism using two fully connected layers with ReLU activation, which learns to assign importance weights to different taxonomic features. The second pathway, termed the abundance network, processes the raw abundance values through a series of transformations, including linear layers, ReLU activation, batch normalization for stable training, and dropout (p=0.3) for regularization. Similarly, the third pathway, the interaction network, maintains an identical structure to the abundance network but processes the data independently to capture potential inter-species interactions. The outputs from the abundance and interaction networks are concatenated and processed through final layers that gradually reduce dimensionality while maintaining the same regularization techniques (batch normalization and dropout). All hidden layers maintain a consistent dimensionality, while the final output layer produces predictions appropriate for the target variable. This architecture enables the model to simultaneously consider both direct abundance effects and potential ecological interactions while automatically learning which features are most relevant for the prediction task.

Figure 3: The FFNN architecture. A flowchart of its components and mathematical functions is utilized to construct the graphs and train the model.

Evaluation Metrics:

To evaluate the performance of the model, various metrics were utilised to assess the performance of the model. The metrics were accuracy, precision, recall, F1 Score, confusion matrices, specificity, and sensitivity. All metrics were calculated using weighted averages to account for potential class imbalance in the dataset. Permutation importance was computed separately for each symptom model (severity and frequency), allowing symptom-specific feature relevance to be evaluated.

Feature Importance Analysis:

For feature importance analysis, the following methods were employed:

1. Permutation Importance: Feature importance was measured by randomly shuffling feature values and observing the decrease in the model's performance. A larger decrease in performance indicates higher feature importance, as it suggests the model heavily relies on that feature for accurate predictions.

2. Feature Correlation: Spearman's correlation between each microbial abundance was applied, providing a measure of the statistical relationship between individual features and the target variable, along with a direction of that measure.

Permutation importance and feature correlation were computed separately for each symptom model (severity and frequency), allowing symptom-specific feature relevance to be evaluated.

Cross Validation:

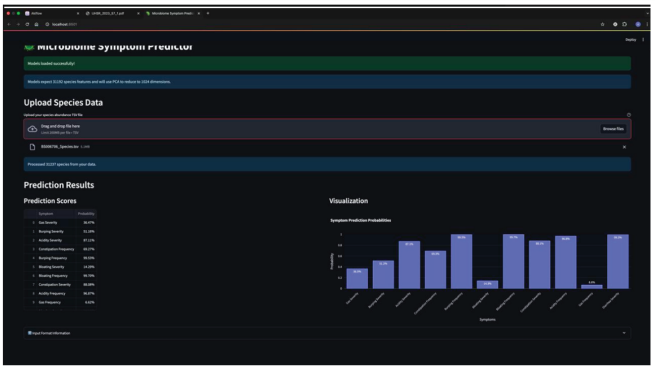
To ensure robust model evaluation, k-fold cross-validation was implemented on the best model (k=5), where the dataset was partitioned into k equal-sized segments. This approach iteratively used k-1 segments for training while reserving one segment for validation, rotating through all possible combinations. This methodology provided a more reliable estimate of the model's generalization performance compared to a single train-test split. The dataset was also tested on 5 external GMTs, and it was able to deliver predictions with an accuracy of ~0.8.

Web Application:

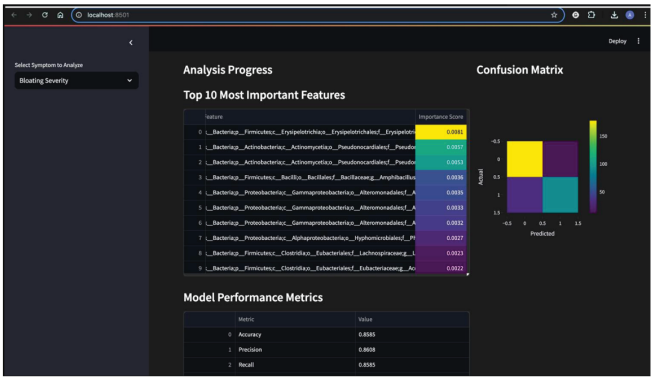
The web applications were then created and deployed. The web applications were built using the Hugging Face platform and are hosted on a HIPAA-compliant cloud server with secure, encrypted data handling. Two web apps were built to include two functionalities: the users can upload their GMT, and the model diagnoses them with IBS symptoms, and clinicians can upload entire datasets on which the model is run, returning validation metrics as well as potential treatment targets (**Figure 4**). Five practicing gastroenterologists reviewed and provided feedback on the web application interfaces and functionality prior to deployment. The web applications are hosted on a HIPAA-compliant cloud with encrypted data transmission. The links of the two web applications are as follows:

1. https://huggingface.co/spaces/anavgupta/ibs-predictions
2. https://huggingface.co/spaces/anavgupta/ibs-dataset

DOI: 10.36838/v7i8.10



(A)



(B)

Figure 4: A layout of web applications. (A). This shows the web application interface developed for Clinicians (B). This shows the web application interface developed for patients.

Result and Discussion

This section presents the results obtained from machine learning and deep learning models applied to the gut microbiome for predicting IBS. The evaluation metrics achieved for FFNNs, RF, AB, and GNNs are listed in Tables 1, 2, 3, and 4, respectively. After cleaning and preprocessing, 1089 samples were retained for analysis. Principal Component Analysis (PCA) reduced dimensionality from 31,259 to 1024. First, RF and AB models were trained, delivering accuracies of 65-75%. After this, GNNs were trained, delivering a 10-15% improvement with accuracies ranging from 80-85%. However, due to its complexity and overfitting on the training set, a lighter FFNN architecture was trained, which not only eliminated complexity but also improved accuracy by ~5%.

Table 1: Evaluation metrics for FFNN models for different symptoms. All symptom severities and frequencies were tested for accuracy, precision, recall, F1-score, specificity, and sensitivity.

FEED FORWARD NEURAL NETWORKS												
Symptom	SEVERITY						FREQUENCY					
	Accuracy	Precision	Recall	F1-Score	Specificity	Sensitivity	Accuracy	Precision	Recall	F1-Score	Specificity	Sensitivity
Bloating	0.88	0.88	0.88	0.88	0.87	0.87	0.86	0.86	0.86	0.86	0.85	0.85
Constipation	0.85	0.85	0.85	0.85	0.76	0.76	0.80	0.80	0.80	0.80	0.74	0.74
Acidity	0.84	0.84	0.84	0.83	0.78	0.78	0.81	0.81	0.8	0.81	0.71	0.71
Diarrhea	0.84	0.83	0.84	0.83	0.77	0.77	0.81	0.82	0.82	0.81	0.82	0.82
Burping	0.90	0.90	0.90	0.90	0.86	0.86	0.84	0.84	0.84	0.84	0.80	0.80

Table 2: Evaluation metrics for RF models for different symptoms. All symptom severities and frequencies were tested for accuracy, precision, recall, F1-score, specificity, and sensitivity.

RANDOM FOREST												
Symptom	SEVERITY						FREQUENCY					
	Accuracy	Precision	Recall	F1-Score	Specificity	Sensitivity	Accuracy	Precision	Recall	F1-Score	Specificity	Sensitivity
Bloating	0.75	0.69	0.59	0.59	0.84	0.63	0.64	0.66	0.68	0.67	0.82	0.66
Constipation	0.78	0.63	0.65	0.64	0.64	0.70	0.61	0.63	0.70	0.67	0.66	0.71
Acidity	0.70	0.66	0.70	0.68	0.65	0.67	0.53	0.50	0.51	0.52	0.65	0.67
Diarrhea	0.68	0.63	0.64	0.67	0.71	0.68	0.58	0.60	0.53	0.52	0.75	0.68
Burping	0.74	0.63	0.62	0.65	0.68	0.72	0.50	0.60	0.54	0.56	0.69	0.72

Table 3: Evaluation metrics for RF models for different symptoms. All symptom severities and frequencies were tested for accuracy, precision, recall, F1-score, specificity, and sensitivity.

ADAPTIVE BOOSTING												
Symptom	SEVERITY						FREQUENCY					
	Accuracy	Precision	Recall	F1-Score	Specificity	Sensitivity	Accuracy	Precision	Recall	F1-Score	Specificity	Sensitivity
Bloating	0.74	0.63	0.60	0.54	0.60	0.66	0.66	0.66	0.68	0.67	0.61	0.66
Constipation	0.74	0.62	0.65	0.63	0.66	0.63	0.67	0.63	0.7	0.67	0.65	0.64
Acidity	0.71	0.64	0.70	0.64	0.65	0.67	0.54	0.50	0.51	0.52	0.65	0.63
Diarrhea	0.61	0.63	0.66	0.67	0.64	0.68	0.55	0.60	0.53	0.52	0.66	0.66
Burping	0.64	0.61	0.68	0.66	0.65	0.63	0.56	0.60	0.54	0.56	0.65	0.63

Table 4: Evaluation metrics for GNN models for different symptoms. All symptom severities and frequencies were tested for accuracy, precision, recall, F1-score, specificity, and sensitivity.

GRAPH NEURAL NETWORKS												
Symptom	SEVERITY						FREQUENCY					
	Accuracy	Precision	Recall	F1-Score	Specificity	Sensitivity	Accuracy	Precision	Recall	F1-Score	Specificity	Sensitivity
Bloating	0.83	0.82	0.81	0.83	0.80	0.76	0.81	0.81	0.81	0.81	0.81	0.80
Constipation	0.81	0.82	0.82	0.82	0.76	0.76	0.75	0.75	0.75	0.75	0.74	0.74
Acidity	0.78	0.79	0.79	0.83	0.74	0.73	0.81	0.76	0.76	0.78	0.72	0.72
Diarrhea	0.84	0.83	0.84	0.83	0.77	0.77	0.81	0.73	0.73	0.73	0.73	0.73
Burping	0.87	0.87	0.86	0.86	0.76	0.76	0.82	0.82	0.82	0.82	0.80	0.80

FFNNs displayed the highest accuracy of 80-85% for symptom frequency and 85-90% for frequencies. Amongst the models built on the severity of different symptoms of IBS, the FFNN model for the severity of burping achieved the highest severity accuracy of 0.90, and amongst the models built on the frequency of different symptoms of IBS, the FFNN model for the frequency of bloating had the highest frequency accuracy of 0.86. ROC-AUC curves were computed for the FFNN, as shown in Figure 5.

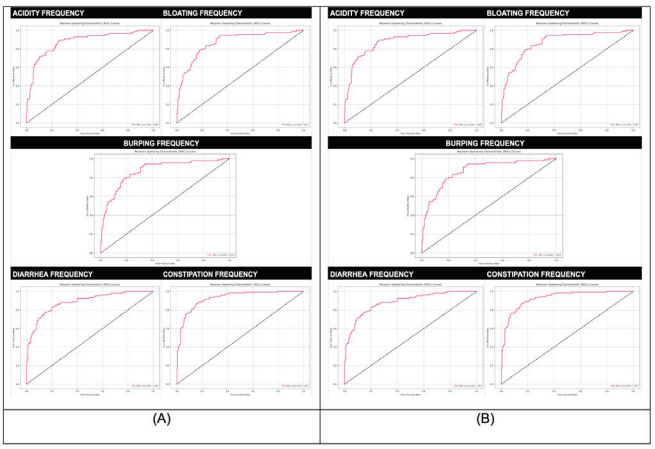


Figure 5: ROC-AUC curves generated for FFNNs using matplotlib. (A) This figure shows the ROC-AUC curves for the models generated for severities of all the symptoms. (B) This figure shows the ROC-AUC curves for the models generated for frequencies of all the symptoms.

To evaluate the model's performance, confusion matrices were built, which gave the frequency of true positives, true neg-

atives, false positives, and false negatives. Confusion matrices give insights into the biases of an ML model and check if it is correctly able to predict high and low outputs. The confusion matrices for different FFNN models built on the frequency and severity of each symptom were computed as listed in Table 5.

Table 5: Confusion Matrices for each symptom severity and frequency. Accuracies for different models lie between 80% and 90%.

FREQUENCIES			SEVERITIES		
ACIDITY			ACIDITY		
	Predicted 0	Predicted 1		Predicted 0	Predicted 1
Actual 0	227	25	Actual 0	213	15
Actual 1	47	44	Actual 1	43	65
BLOATING			BLOATING		
	Predicted 0	Predicted 1		Predicted 0	Predicted 1
Actual 0	165	19	Actual 0	172	17
Actual 1	36	88	Actual 1	25	97
BURPING			BURPING		
	Predicted 0	Predicted 1		Predicted 0	Predicted 1
Actual 0	121	30	Actual 0	145	24
Actual 1	36	105	Actual 1	34	98
CONSTIPATION			CONSTIPATION		
	Predicted 0	Predicted 1		Predicted 0	Predicted 1
Actual 0	207	22	Actual 0	232	20
Actual 1	35	69	Actual 1	41	49
DIARRHEA			DIARRHEA		
	Predicted 0	Predicted 1		Predicted 0	Predicted 1
Actual 0	202	20	Actual 0	263	11
Actual 1	35	71	Actual 1	47	48

A validation was performed on all the FFNN models, after which the model displayed strong validation results with a 1-2% difference in frequencies and a 2-4% difference in severities, as shown in Table 6. The model demonstrated consistent performance across all folds, with minimal variance between different data splits.

Table 6: Fivefold validation results. The metrics tested were accuracy, precision, recall, and F1 score.

CROSS VALIDATION								
Symptom	SEVERITY				FREQUENCY			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
Bloating	± 0.0211	± 0.0215	± 0.0233	± 0.0211	± 0.0159	± 0.0156	± 0.0159	± 0.0166
Constipation	± 0.0241	± 0.0244	± 0.0241	± 0.0241	± 0.0171	± 0.0198	± 0.0171	± 0.0185
Acidity	± 0.0164	± 0.0128	± 0.0141	± 0.0128	± 0.0191	± 0.0213	± 0.0191	± 0.0207
Diarrhea	± 0.0408	± 0.0379	± 0.0334	± 0.0379	± 0.0089	± 0.0091	± 0.0089	± 0.0085
Burping	± 0.0285	± 0.0290	± 0.0303	± 0.0290	± 0.0186	± 0.0193	± 0.0186	± 0.0195

Permutation importance and feature correlation computed separately for each symptom model allowed symptom-specific feature relevance to be evaluated. This approach provided complementary insights into both the predictive power of each taxon and the direction of the relationship between the taxon and the output. Through this, a clinician will be able to identify whether a particular taxon is negatively or positively correlated with the output and how strongly it impacts the model's performance. The top 10 taxa for each symptom, namely diarrhoea, constipation, burping, acidity, and bloating, along with their importance and correlation to IBS, are shown in Tables 7, 8, 9, 10, and 11, respectively.

Table 7: The top ten most important taxa for diarrhea. Arranged from most influential to least influential according to their permutation scores. Correlation coefficients are also included.

DIARRHEA FREQUENCY			DIARRHEA SEVERITY		
TAXON	IMPORTANCE	CORRELATION	TAXON	IMPORTANCE	CORRELATION
Amycolatopsis granulosa	0.0065	-0.0145	*Amycolatopsis granulosa*	0.0087	0.0176
[Clostridium] cocleatum	0.0059	0.0278	*Aeromonas veronii*	0.004	0.0042
Acidaminococcus fermentans	0.0054	-0.0432	*Amniculibacterium sp. G2-70*	0.0037	-0.0185
Acinetobacter bereziniae	0.0051	-0.0089	*Alteromonas sp. RKMC-009*	0.0036	0.0396
Anaerobutyricum soehngenii	0.0047	-0.0256	*[Clostridium] cocleatum*	0.0036	-0.0042
[Enterobacter] lignolyticus	0.0045	0.0183	*[Clostridium] populeti*	0.0036	-0.0122
Alcanivorax profundi	0.0043	-0.0521	*Acinetobacter junii*	0.0036	0.0079
Acetivibrio clariflavus	0.0041	-0.0167	*Acholeplasma oculi*	0.0035	0.0384
Anaerococcus provencensis	0.004	0.0092	*Acetivibrio mesophilus*	0.0032	0.0272
Amniculibacterium sp. G2-70	0.0038	-0.0394	*Acidaminococcus fermentans*	0.0032	-0.0335

Table 8: The top ten most important taxa for constipation. Arranged from most influential to least influential according to their permutation scores. Correlation coefficients are also included.

CONSTIPATION FREQUENCY			CONSTIPATION SEVERITY		
TAXON	IMPORTANCE	CORRELATION	TAXON	IMPORTANCE	CORRELATION
Amycolatopsis circi	0.0064	0.0315	*Amycolatopsis granulosa*	0.0082	0.036
Aminicella lysinilytica	0.0041	-0.0123	*Acetitomaculum ruminis*	0.0053	0.2145
Acidaminococcus fermentans	0.0027	-0.0457	*Anaerophilus nitritogenes*	0.0043	-0.0497
Amphritea atlantica	0.0026	0.0115	*Acetivibrio saccincola*	0.0042	0.0202
Aminomonas paucivorans	0.0025	0.034	*Amycolatopsis sp. CA-128772*	0.0034	-0.0072
Lachnoclostridium aminophilum	0.0023	-0.0122	*Alkalicoccus urumqiensis*	0.0032	0.0071
Eubacterium cellulosolvens	0.0023	-0.0181	*Acinetobacter guillouiae*	0.0029	-0.0356
Aliarcobacter cryaerophilus	0.0023	0.021	*[Clostridium] dakarense*	0.0023	0.0685
Achromobacter sp. GG226	0.0022	0.0465	*Acholeplasma oculi*	0.0023	-0.064
Akkermansia sp. BIOML-A61	0.002	0.0112	*Acinetobacter seiferti*	0.0023	0.0661

Table 9: The top ten most important taxa for burping. Arranged from most influential to least influential according to their permutation scores. Correlation coefficients are also included.

BURPING FREQUENCY			BURPING SEVERITY		
TAXON	IMPORTANCE	CORRELATION	TAXON	IMPORTANCE	CORRELATION
Achromobacter spanius	0.0053	0.0458	*Amycolatopsis granulosa*	0.0075	-0.0126
Amycolatopsis taiwanensis	0.0053	-0.0141	*Amniculibacterium sp. G2-70*	0.0033	-0.0325
Anaerofilum sp. BX8	0.0049	-0.0199	*[Clostridium] cocleatum*	0.0031	0.0299
Aliarcobacter lanthieri	0.004	-0.0375	*[Enterobacter] lignolyticus*	-0.0028	-0.0298
Aeromicrobium sp. zg-Y50	0.0037	0.0132	*Acetivibrio clariflavus*	-0.0027	0.0436
Amycolatopsis granulosa	0.0034	0.0045	*Eubacterium cellulosolvens*	0.0027	-0.163
[Clostridium] populeti	0.0033	-0.0274	*[Clostridium] polysaccharolyticum*	0.0025	0.0261
Alcaligenes sp. SORT26	0.0033	-0.0235	*Acidovorax radicis*	0.0024	0.0032
Acidaminococcus fermentans	0.0029	-0.0307	*Alcanivorax profundi*	0.0023	0.0019
Acinetobacter seiferti	0.0028	0.3497	*Agathobaculum sp. NSJ-28*	0.0022	-0.0081

DOI: 10.36838/v7i8.10

Table 10: The top ten most important taxa for acidity. Arranged from most influential to least influential according to their permutation scores. Correlation coefficients are also included.

ACIDITY FREQUENCY			ACIDITY SEVERITY		
TAXON	IMPORTANCE	CORRELATION	TAXON	IMPORTANCE	CORRELATION
[Enterobacter] lignolyticus	-0.0026	-0.0359	*Amycolatopsis circi*	0.0068	-0.0009
[Flexibacter] sp. ATCC 35103	-0.0021	0.0043	*Anaerobranca californiensis*	0.0057	-0.0198
Acinetobacter bereziniae	0.002	0.0446	*Amphritea atlantica*	0.0053	-0.0117
Acetivibrio ethanolgignens	-0.002	0.0028	*Acetobacter pasteurianus*	0.0044	0.0207
Acetitomaculum ruminis	-0.0019	-0.0118	*Amniculibacterium sp. G2-70*	0.0043	-0.0362
Anaerobutyricum soehngenii	-0.0019	-0.0189	*Romboutsia dakarense*	0.0041	-0.0094
Acetanaerobacterium sp. MSJ-12	-0.0018	-0.0333	*Acinetobacter sp. 105-3*	0.0041	-0.0624
[Clostridium] methoxybenzovorans	-0.0018	-0.041	*Aminicella lysinilytica*	0.0041	-0.0012
Acidaminococcus massiliensis	-0.0018	0.0158	*Alcanivorax gelatiniphagus*	0.0038	-0.0023
Alkalihalobacterium elongatum	-0.0018	0.0014	*Amycolatopsis sp. AA4*	0.0038	-0.0568

Table 11: The top ten most important taxa for bloating. Arranged from most influential to least influential according to their permutation scores. Correlation coefficients are also included.

BLOATING FREQUENCY			BLOATING SEVERITY		
TAXON	IMPORTANCE	CORRELATION	TAXON	IMPORTANCE	CORRELATION
Anabaena lutea	0.0117	-0.0012	*Amycolatopsis granulosa*	0.0089	0.0226
[Ruminococcus] lactaris	0.0092	-0.0569	*Amniculibacterium sp. G2-70*	0.005	-0.0175
Anaerococcus provencensis	0.0065	-0.0793	*Megasphaera vaginalis*	0.0042	0.0336
Acanthamoeba castellanii	0.0057	0.0401	*Amphritea atlantica*	0.0039	0.0188
Achromobacter spanius	0.0051	0.057	*Amycolatopsis circi*	0.0039	-0.0316
Acetivibrio clariflavus	0.0051	-0.0246	*Erysipelatoclostridium cocleatum*	0.0037	0.0345
Anaerococcus sp. mt242	0.0049	-0.0039	*Anaerococcus lactolyticus*	0.0033	-0.0041
[Eubacterium] cellulosolvens	0.0047	-0.0191	*Aminicella lysinilytica*	0.0032	-0.0033
Alteromonas sp. RKMC-009	0.0047	0.0232	*Amycolatopsis niigatensis*	0.0032	0.0624
Abyssibacter profundi	0.0044	-0.0353	*Amycolatopsis sp. AA4*	0.0032	-0.0344

Conclusion

This study developed and validated a novel deep learning approach for IBS diagnosis using gut microbiome data from 1,089 Indian patients. The FFNN achieved 85-90% accuracy for symptom severities and 80-85% for frequencies, outperforming existing models by 15-20%. 71 unique taxonomic biomarkers were identified, with Amycolatopsis, Acinetobacter, Clostridium, and Acetivibrio species occurring repeatedly across symptoms and having the highest permutation scores. The model's performance was validated through 5-fold cross-validation and external testing, demonstrating robust generalization. Two web applications were developed for clinical use, enabling both individual diagnosis and dataset-wide analysis. This work represents a significant advancement in objective IBS diagnosis, particularly for the understudied South Asian population, while providing specific microbial targets for therapeutic intervention. Future work should incorporate longitudinal data and additional health metrics to further improve diagnostic accuracy. The identified biomarkers require experimental validation to confirm their biological relevance in IBS pathophysiology.

Acknowledgments

AG would like to thank Max Kushnir (Co-Founder and CSO, Sova Health) for providing the dataset for the research project. AG would also like to thank Mr. Pankaj Aggarwal (Data Scientist, Duke University) for his valuable guidance and feedback. AG would like to acknowledge Aashna Saraf, Founder of CreatED, for providing valuable feedback and guidance throughout the project.

References

1. MayoClinic *Irritable Bowel Syndrome*; USA, 2024.
2. Oka, P.; Parr, H.; Barberio, B.; Black, C. J.; Savarino, E. V.; Ford, A. C., Global prevalence of irritable bowel syndrome according to Rome III or IV criteria: a systematic review and meta-analysis. *Lancet Gastroenterol Hepatol* **2020**, *5* (10), 908-917.
3. Menees, S.; Chey, W., The gut microbiome and irritable bowel syndrome. *F1000Res* **2018**, *7*.
4. Zhao, Y.; Zhu, S.; Dong, Y.; Xie, T.; Chai, Z.; Gao, X.; Dai, Y.; Wang, X., The Role of Gut Microbiome in Irritable Bowel Syndrome: Implications for Clinical Therapeutics. *Biomolecules* **2024**, *14* (12), 1643.
5. Camilleri, M., Diagnosis and Treatment of Irritable Bowel Syndrome: A Review. *Jama* **2021**, *325* (9), 865-877.
6. Auslander, N.; Gussow, A. B.; Koonin, E. V., Incorporating Machine Learning into Established Bioinformatics Frameworks. *Int J Mol Sci* **2021**, *22* (6).
7. Fukui, H.; Nishida, A.; Matsuda, S.; Kira, F.; Watanabe, S.; Kuriyama, M.; Kawakami, K.; Aikawa, Y.; Oda, N.; Arai, K.; Matsunaga, A.; Nonaka, M.; Nakai, K.; Shinmura, W.; Matsumoto, M.; Morishita, S.; Takeda, A. K.; Miwa, H., Usefulness of Machine Learning-Based Gut Microbiome Analysis for Identifying Patients with Irritable Bowels Syndrome. *J Clin Med* **2020**, *9* (8).
8. Hernández Medina, R.; Kutuzova, S.; Nielsen, K. N.; Johansen, J.; Hansen, L. H.; Nielsen, M.; Rasmussen, S., Machine learning and deep learning applications in microbiome research. *ISME Communications* **2022**, *2* (1), 98.
9. McDonald, D.; Hyde, E.; Debelius, J. W.; Morton, J. T.; Gonzalez, A.; Ackermann, G.; Aksenov, A. A.; Behsaz, B.; Brennan, C.; Chen, Y.; DeRight Goldasich, L.; Dorrestein, P. C.; Dunn, R. R.; Fahimipour, A. K.; Gaffney, J.; Gilbert, J. A.; Gogul, G.; Green, J. L.; Hugenholtz, P.; Humphrey, G.; Huttenhower, C.; Jackson, M. A.; Janssen, S.; Jeste, D. V.; Jiang, L.; Kelley, S. T.; Knights, D.; Kosciolek, T.; Ladau, J.; Leach, J.; Marotz, C.; Meleshko, D.; Melnik, A. V.; Metcalf, J. L.; Mohimani, H.; Montassier, E.; Navas-Molina, J.; Nguyen, T. T.; Peddada, S.; Pevzner, P.; Pollard, K. S.; Rahnavard, G.; Robbins-Pianka, A.; Sangwan, N.; Shorenstein, J.; Smarr, L.; Song, S. J.; Spector, T.; Swafford, A. D.; Thackray, V. G.; Thompson, L. R.; Tripathi, A.; Vázquez-Baeza, Y.; Vrbanac, A.; Wischmeyer, P.; Wolfe, E.; Zhu, Q.; Knight, R., American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems* **2018**, *3* (3).
10. Kanz, C.; Aldebert, P.; Althorpe, N.; Baker, W.; Baldwin, A.; Bates, K.; Browne, P.; van den Broek, A.; Castro, M.; Cochrane, G.; Duggan, K.; Eberhardt, R.; Faruque, N.; Gamble, J.; Diez, F. G.; Harte, N.; Kulikova, T.; Lin, Q.; Lombard, V.; Lopez, R.; Mancuso, R.; McHale, M.; Nardone, F.; Silventoinen, V.; Sobhany, S.; Stoehr, P.; Tuli, M. A.; Tzouvara, K.; Vaughan, R.; Wu, D.; Zhu, W.; Apweiler, R., The EMBL Nucleotide Sequence Database. *Nucleic Acids Res* **2005**, *33* (Database issue), D29-33.
11. Kim, J. H.; Lin, E.; Pimentel, M., Biomarkers of Irritable Bowel Syndrome. *J Neurogastroenterol Motil* **2017**, *23* (1), 20-26.
12. Ford, A. C.; Bercik, P.; Morgan, D. G.; Bolino, C.; Pintos-Sanchez, M. I.; Moayyedi, P., Validation of the Rome III criteria for the diagnosis of irritable bowel syndrome in secondary care. *Gastroenterology* **2013**, *145* (6), 1262-70.e1.

Authors

Anav Gupta is presently a senior at The Shri Ram School, Moulsari, Gurugram, Haryana, India. He is keenly interested in technology, Computer Science, and Math. He is a passionate photographer and serves as the editor-in-chief for The Shri Ram School magazine. He has won several prestigious STEM Awards, including being selected as a finalist in the IRIS National Science Fair, and presented this work in the Computational Biology and Bioinformatics Category.

Nirupma Singh is a Bioinformatics Scientist with a doctorate in Biotechnology and Bioinformatics from the University of Delhi, with six years of hands-on research and development experience. Her journey is marked by a robust foundation in Machine Learning and Python, with five years of expertise. Proficient in Linux and cloud servers like AWS. She excels in structural biology, systems biology, protein/gene network analysis, data mining, and computational genomics.

DOI: 10.36838/v7i8.10

IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Developing Biodegradable Nanoparticles from Corn for Treating Brain Cancer: Insights from Live Cell Imaging

Donghyeon Oh

Avon Old Farms School, 500 Old Farms Rd, Avon, CT 06001, USA; gosyber1@gmail.com

ABSTRACT: Nanoparticles (NPs) have emerged as the most suitable means of delivering drugs to target cells. This specific functional property of NPs can extend conventional cancer treatments' current coverage. Corn starch-based nanoparticles are a possibility for cancer treatment. Because of the NPs' biocompatibility, low toxicity, and friendly nature, several drugs based on NPs have been synthesized. The purpose of this research is to determine the ability of corn-derived nanoparticles (cNPs) to treat human brain cancers, including glioblastoma and neuroblastoma. This research mainly concentrates on how these cNPs can exploit the properties of cancer in corn to locate and destroy cancer cells. This study seeks to present a method in cancer therapy that is effective and in line with sustainability and accessibility principles through an extraction process that isolates these nanoparticles and tests them on cancer cell cultures. The findings suggesting a decrease in cancer cell density after exposure to cNP are a step forward in applying plant-derived nanoparticles in medical oncology.

KEYWORDS: Nanoparticles, Corn, Brain Cancer, Cell Viability, Live Cell Imaging.

Introduction

Nanoparticles, or NPs, contain 1 to 100 nm parameters. The size of NPs may affect characteristics such as ductility, rigidity, and melting points that differ from those of their larger-sized counterparts.[1] These NPs can be easily engineered for drug delivery systems and exhibit specific characteristics that render them valuable across numerous domains.[2] Therefore, research on NPs is rapidly increasing in medicine, cosmetics, sports, and aerospace engineering. In particular, they carry chemotherapeutic drugs to cancer cells.

Edible nanoparticles (ENPs) refer to nano-sized vesicles from edible plants containing plant-derived microRNAs, proteins, lipids, and phytochemicals.[3] The enzymic substances are extracted from different plant species like ginger, lemon, tomato, broccoli, orange, kiwi, pear, soybean, grapefruit, and coconut.[4] Lately, it has been observed that nanoscale particles extracted from corn (CNP) grain have demonstrated anti-tumor properties.[5] The anticancer effects of CNPs are attributed to the presence of vitamins, minerals, and xanthophylls found in corn, all of which exhibit potent anticancer properties.[6] The simplicity of corn nanoparticles synthesis, which can be gained from maize plants, a high-productivity harvest that can be raised speedily and in huge volumes, makes them an affordable option for managing carcinomas.[7] The ultimate goal of engineering edible nanoparticles is to develop them as a promising avenue for cancer treatment.

Current therapeutic options for brain cancer are often constrained by their ineffectiveness: the adverse side effects and difficulty in delivering treatments across the blood-brain barrier.[8] Traditional approaches like surgical interventions, radiation therapy, and chemotherapy are still used for brain cancer treatment.[9] Furthermore, the invasive nature of surgeries and the considerable systemic toxicity from chemotherapy are still considered significant side effects of brain cancer treatment.[10]

This study analyzed whether corn nanoparticles have an anticancer effect on treating human brain cancer. The CNPs were isolated and tested to see if they could induce cancer cell death. The two most common types of brain cancer cell lines were used to analyze the effect of cNPs on brain cancer cell death: glioblastoma and neuroblastoma. Then, explore the impact of cNPs that could slow down the growth of brain cancer cells. This new research might give us a better way to use nanoparticles to treat brain cancer, making treatments safer and more effective.

Methods

Isolation of Corn-Derived Nanoparticles:

Figure 1: Workflow for cNP isolation. Corn kernels were homogenized, centrifuged, filtered, and subjected to sucrose gradient ultracentrifugation to extract the corn-derived nanoparticle (cNP) layer. This process yields a reproducible and pure cNP sample suitable for biological testing.

The corn was bought from a market and washed three times with distilled water. 100 g of corn kernels were mixed with 100 mL of distilled water and homogenized for 5 minutes using a food processor. The homogenized corn juice was centrifuged at

DOI: 10.36838/v7i8.11

2000 × g for 30 minutes. Then, the samples were centrifuged at 5000 × g for 40 min. Lastly, the samples were centrifuged at 10,000 × g for 1 hour at a temperature of 4˚C. Afterward, the supernatant was filtered using a 0.45 µm-pore size syringe filter made by Millipore. To the filtrate (38 mL), 2 mL of a 60% sucrose solution was added in a 50 mL ultracentrifuge tube. The mixture was then ultra-centrifuged at 100,000 × g for 160 minutes at 4˚C using a Beckman Optima XL-100 K with an SW28 centrifuge rotor made by Beckman Coulter. The top supernatant was removed with a syringe. The yellow-colored cNP layer above the clear 60% sucrose solution was collected carefully with a syringe. The isolated cNPs were stored at -80˚C. The isolated cNPs from 100 g of corn were indicated as 100% concentration (Figure 1).

Cell Culture and Treatment Conditions:

The initial cell number of 0.12 x 106 cells was prepared in a 12-well culture plate for both A172 and SH-SY5Y cells. Then, various concentrations (0, 2.5, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50%) of cNPs were added to each cell culture well. The cell cultures mixed with cNPs were incubated for seven days. Then, the cells were photographed using an EVOS M5000 microscope (Invitrogen). The cells were visualized and photographed using transmitted light and a phase objective with 10X magnification. The selected concentration range of 0–50% cNP was chosen based on preliminary trials and existing literature on plant-derived nanoparticles, allowing the investigation of both low-dose cellular responses and high-dose cytotoxic effects. This range aimed to identify the minimum effective dose capable of inducing morphological changes and cell death in brain cancer cells.

Cell Viability Assay:

To quantify cell viability, LUNA-FL™ dual fluorescence cell counter was used to quantify cell viability. After the cells were stained with acridine orange (AO) and propidium iodide (PI), live and dead cells were analyzed. AO permeates all cells and emits a green fluorescence, marking both live and dead cells, whereas PI penetrates only dead cells, emitting a red fluorescence.

Result and Discussion

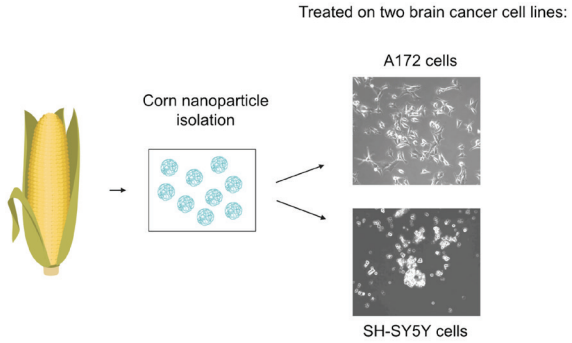


Figure 2: Experimental overview. Schematic illustrating the isolation of cNPs and their application to glioblastoma (A172) and neuroblastoma (SH-SY5Y) cell lines for viability assessment. The workflow enables systematic evaluation of plant-derived nanoparticles as anti-cancer agents in brain cancer models.

We aimed to investigate the effectiveness of corn nanoparticles in treating human brain cancer. Firstly, we isolated nanoparticles from corn (Figure 2). Since brain cancer requires novel and effective therapies, we investigated the two most common human brain cancer types: glioblastoma and neuroblastoma. Therefore, we tested the potential anticancer effects on two cancer cell lines, A172 and SH-SY5Y cells (Figure 2). Also, we aimed to determine the therapeutic potential of corn-derived nanoparticles by testing on the growth and proliferation of brain cancer cells.

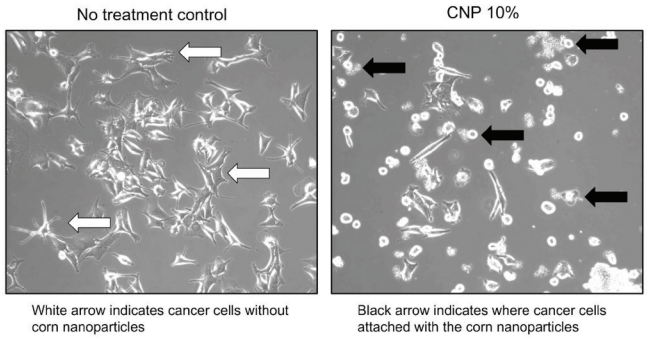


Figure 3: Morphological changes in A172 glioblastoma cells after cNP exposure (10%). Black arrows indicate cells with attached cNPs and rounded morphology; white arrows indicate untreated cells with normal morphology. Exposure to cNPs leads to loss of adhesion and altered morphology in cancer cells, suggesting early signs of cytotoxicity.

Our current research is focused on examining the effects of corn nanoparticles (CNP) on the morphology and behavior of glioblastoma cancer cells. In Figure 3, the visual representations of cancer cell morphology in both the untreated control and CNP10% samples are different. Typically, A172 cancer cells exhibit a distinct glia-like, star-like structure, depicted in the images of the untreated control (Figure 3 left). However, after exposure to CNP10%, the cancer cells underwent a remarkable transformation, adopting a rounded shape. This indicates the loss of their ability to adhere to the surface of the culture plate. Also, the presence of CNPs was visually apparent as they formed sizable and firmly attached complexes on the surface of the cancer cells (Figure 3, right). Even after subjecting the cells to agitation, the corn nanoparticles remained steadfastly bound to the cells. This result indicates that CNP showed a profound and enduring interaction.

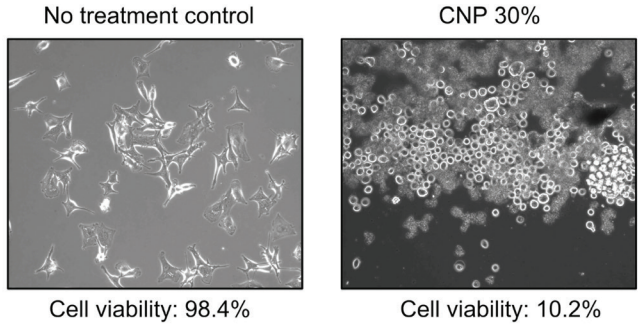


Figure 4: Cell viability of A172 cells after 30% cNP treatment. Significant reduction in viable cell count indicates cytotoxic effects of high-dose cNP exposure. At 30% concentration, cNPs reduce glioblastoma cell survival to 10.2%, demonstrating potent anticancer activity.

Next, we experimented by increasing the CNP concentration to 30% and checking how well the cancer cells survived. Not surprisingly, the control group with no treatment had a cell survival rate of 98.4%. However, when we hit the A172 cells with 30% CNP, their survival rate plummeted to just 10.2% (Figure 4). This tells us that the round shape of the A172 cells we saw in Figure 6 was caused by the CNP treatment-induced cell death. Therefore, this result indicates that CNP treatment showed promising results in treating A172 glioblastoma cancer cells.

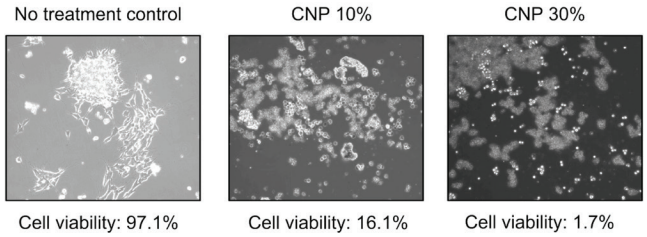


Figure 5: Viability of SH-SY5Y cells under 10% and 30% cNP treatments. Higher cNP concentrations correlate with increased cancer cell death. cNPs induce strong dose-dependent cytotoxicity in neuroblastoma cells, reinforcing their therapeutic potential.

In the next experiment, we used the SH-SY5Y neuroblastoma cancer cell line and tested the effect of CNPs on inducing cancer cell death. We tested three conditions: no treatment, CNP 10%, and CNP 30%. Then, cell viability was analyzed under each condition. Like with the A172 cancer cell line, CNP caused the neuroblastoma cancer cells to induce cell death (Figure 5). Specifically, CNP 10% reduced the cell viability to 16.1%, while CNP 30% lowered it to 1.7%. This indicates that higher concentrations of CNP result in more cancer cell death.

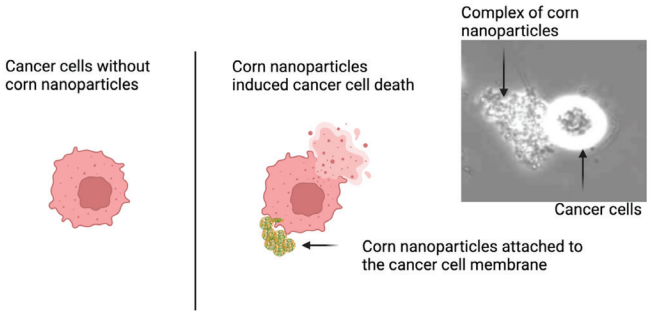


Figure 6: Proposed mechanism of cNP-induced cell death. Illustration summarizing cNP interaction with brain cancer cells, promoting adhesion, morphological changes, and apoptosis. cNPs act through physical binding and membrane disruption to induce cancer cell death, supporting their role as plant-derived nanomedicine agents.

From the current study, we can assertively confirm that CNPs possess potent anticancer properties against glioblastoma and neuroblastoma cell lines. This may offer a new method of more effective and safer treatments for brain cancer. Indeed, CNPs that can directly bind to brain cancer cells and cause cell death may serve as a potential therapeutic approach (Figure 6). This work indicates the possibility of using plant-derived nanoparticles in oncology. This result also shows the need to continue finding environmentally friendly and sustainable sources of agents in the discovery of new cancer therapies.

Nevertheless, several limitations should be noted regarding the outcomes presented in this study. Since all experiments were conducted in vitro, future research should involve in vivo validation using animal models of glioblastoma and neuroblastoma to assess the biodistribution, blood-brain barrier penetration, and systemic toxicity of corn-derived nanoparticles. Additionally, mechanistic studies using molecular assays such as caspase activation, ROS quantification, or gene expression profiling are needed to elucidate the apoptotic or necrotic pathways involved in cNP-induced cancer cell death. Finally, comparing the efficacy of cNPs with existing chemotherapeutics could help position these nanoparticles as potential adjuncts or alternatives in current treatment regimens.

The second drawback is that there is no comparison of cNPs with the existing therapies for brain cancer that would give a better idea of the effectiveness of cNPs. Last but not least, the long-term impacts of cNPs on human health and specifically on the blood-brain barrier (BBB) and healthy brain tissue are yet to be explored. Overcoming these limitations will be important for the further development of cNPs in clinical practice and the optimal use of cNPs in the treatment of brain cancer.

cNPs may exert anticancer effects through several potential molecular mechanisms. First, their phytochemical components—such as polyphenols, flavonoids, and xanthophylls—can induce reactive oxygen species (ROS) generation within cancer cells, leading to oxidative stress and mitochondrial dysfunction. This oxidative damage can activate intrinsic apoptotic pathways, including mitochondrial membrane permeabilization and caspase-3/9 activation, ultimately resulting in programmed cell death. Additionally, cNPs may disrupt cancer cell membrane integrity through direct adhesion and surface interaction, impairing adhesion-dependent signaling and promoting detachment-induced apoptosis (anoikis). Some plant-derived nanoparticles have also been shown to modulate cell cycle regulators, such as downregulating cyclin D1 or upregulating p21, thereby inhibiting proliferation. In brain cancer specifically, the ability of cNPs to cross or interact with the BBB—potentially facilitated by their nano-size and surface bioactivity—makes them promising carriers for both intrinsic cytotoxic effects and targeted drug delivery. Further mechanistic studies are required to validate these pathways and clarify the role of specific bioactive molecules present in corn-derived nanoparticles.

■ Conclusion

This study demonstrates that corn-derived nanoparticles (cNPs) exhibit significant anticancer activity against glioblastoma and neuroblastoma cell lines, causing notable morphological changes and a dose-dependent reduction in cell viability. These findings highlight the potential of plant-based nanomedicine in brain cancer therapy. However, as the experiments were conducted entirely *in vitro*, the results may not fully reflect the complex biological environment of human brain tumors. To advance the clinical relevance of cNPs, future studies should involve in vivo models to assess their biodis-

DOI: 10.36838/v7i8.11

tribution, toxicity, and ability to cross the blood-brain barrier. Additionally, mechanistic investigations using molecular assays are needed to clarify the pathways involved in cNP-induced cell death. Comparative studies with existing chemotherapeutic agents would also help contextualize the efficacy of cNPs and support their development as complementary or alternative treatments.

Acknowledgments

I want to thank the IRIS Lab and Prof. Lee for supporting this research project.

References

1. Murthy, S. K. Nanoparticles in Modern Medicine: State of the Art and Future Challenges. *Int. J. Nanomedicine* **2007**, *2* (2), 129–141.
2. Altammar, K. A. A Review on Nanoparticles: Characteristics, Synthesis, Applications, and Challenges. *Front. Microbiol.* **2023**, *14*, 1155622. DOI: 10.3389/fmicb.2023.1155622.
3. Zhang, M.; Viennois, E.; Xu, C.; Merlin, D. Plant Derived Edible Nanoparticles as a New Therapeutic Approach against Diseases. *Tissue Barriers* **2016**, *4* (2), e1134415. DOI: 10.1080/21688370.2015.1134415.
4. Salvia-Trujillo, L.; Soliva-Fortuny, R.; Rojas-Graü, M. A.; McClements, D. J.; Martín-Belloso, O. Edible Nanoemulsions as Carriers of Active Ingredients: A Review. *Annu. Rev. Food Sci. Technol.* **2017**, *8*, 439–466. DOI: 10.1146/annurev-food-030216-025908.
5. Sasaki, D.; Kusamori, K.; Takayama, Y.; Itakura, S.; Todo, H.; Nishikawa, M. Development of Nanoparticles Derived from Corn as Mass Producible Bionanoparticles with Anticancer Activity. *Sci. Rep*. **2021**, *11* (1), 22818. DOI: 10.1038/s41598-021-02241-y.
6. Luo, X.; Wu, S.; Xiao, M.; Gu, H.; Zhang, H.; Chen, J.; Liu, Y.; Zhang, C.; Zhang, J. Advances and Prospects of Prolamine Corn Protein Zein as Promising Multifunctional Drug Delivery System for Cancer Treatment. *Int. J. Nanomedicine* **2023**, *18*, 2589–2621. DOI: 10.2147/IJN.S402891.
7. Sasaki, D.; Kusamori, K.; Nishikawa, M. Delivery of Corn-Derived Nanoparticles with Anticancer Activity to Tumor Tissues by Modification with Polyethylene Glycol for Cancer Therapy. *Pharm. Res.* **2023**, *40* (4), 917–926. DOI: 10.1007/s11095-022-03431-7.
8. Bhowmik, A.; Khan, R.; Ghosh, M. K. Blood Brain Barrier: A Challenge for Effectual Therapy of Brain Tumors. *Biomed Res. Int.* **2015**, *2015*, 320941. DOI: 10.1155/2015/320941.
9. Aquilanti, E.; Wen, P. Y. Current Therapeutic Options for Glioblastoma and Future Perspectives. *Expert Opin. Pharmacother*. **2022**, *23* (14), 1629–1640. DOI: 10.1080/14656566.2022.2125302.
10. Laquintana, V.; Trapani, A.; Denora, N.; Wang, F.; Gallo, J. M.; Trapani, G. New Strategies to Deliver Anticancer Drugs to Brain Tumors. *Expert Opin. Drug Deliv*. **2009**, *6* (10), 1017–1032. DOI: 10.1517/17425240903167942.

Author

Donghyeon Oh is a student at Avon Old Farms School. He is interested in biology and plans to continue his studies in this field. He loves researching and is strongly curious about living cells and cell biology.

IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Urine as a Nitrogen Source for *Lepidium sativum*: Creation of A Novel Synthetic Urine Testing Model and Product

Thomas H. Bill, Leslie B. Yang

Harrisonburg High School, 1001 Garbers Church Rd, Harrisonburg, VA, 22801, USA; tomhenbill@gmail.com
Mentor: Timothy J. Bill

ABSTRACT: The Haber-Bosch Process produces a majority of the world's ammonia-based fertilizers. Due to the inherent explosive nature of ammonium nitrate, the requisite fossil fuels, and the massive carbon dioxide (CO_2) emissions, it is time to evaluate other sources of nitrogen to feed the world's 8 billion people. In many studies, the use of human urine as a fertilizer has been described, but this concept is challenging on many levels. This paper describes a novel method to create, test, and store solid synthetic urine as a fertilizer for *Lepidium sativum* (garden cress). Garden cress fertilized with synthetic urine had a significantly higher yield in mass than the control group. Synthetic urine in its solid, powdered form decreases storage and transportation issues compared with large volumes of liquid urine. Additionally, this study demonstrates that solid urine is shelf-stable due to its bacteriostatic properties and low hygroscopicity. Fertilizing with human urine can provide a renewable resource in agricultural systems, reducing the need for the Haber Bosch Process and its consequences.

KEYWORDS: Plant Sciences, Agriculture and Agronomy, Haber Bosch Process, Sustainable Fertilizer, Synthetic Urine.

Introduction

Using nitrogen-based fertilizers for crops is a tremendous asset, as we must now feed 8 billion people. Most of this nitrogen is captured from the atmosphere and converted into ammonia and solid ammonium nitrate. This energy-intensive chemical reaction, the Haber-Bosch process, demands tremendous amounts of non-renewable energy and generates 1.4% of all global CO_2 emissions.[1] Additionally, ammonium nitrate is highly explosive, and it has been used in acts of terror and has resulted in tragic deaths from inadvertent explosions.[2,3]

Using human waste as a fertilizer is a centuries-old technique; however, the modern use of biosolids and sewage sludge is not without controversy. There are growing concerns about the safety of this modality. These include the risk of viral and bacterial pathogens, heavy metals, and toxic organic compounds such as PCBs, dioxins, and even hospital waste.[4] While permitted in the United States and Europe, despite stringent regulations, the processing cost runs in the billions of dollars, and there are still no guarantees about long-term safety.

On the other hand, the process described in this paper results in solid urine that is more sanitary than sewage sludge (bacteriostatic) and comes without the risk of toxic organic and inorganic compounds. Because of the high urea content, urine is an exemplary nitrogen source. The infrastructure to isolate urine from the waste stream has been developed and studied, and when processed, the potential of diluted urine has been analyzed in some Scandinavian countries.[5] In liquid form, however, urine remains susceptible to microbial growth and presents transportation challenges due to the large volumes of liquid that must be delivered to farms. This paper presents a unique process to create a shelf-stable, solid urine fertilizer that can be stored and prepared at the site and time of application. On an environmental level, a long-term goal would be to scale this process up to capture urine city-wide, creating fertilizer and diverting the nitrogen stream. Mechanical vapor recompression technology exists and could perform large-scale urine dehydration cost-effectively and minimize energy use.[6]

Not only is the creation of solid urine unique, but using updated medical-grade synthetic urine for agricultural testing has not been previously described. The process and testing are outlined using *Lepidium sativum* (garden cress). This research complements other studies that have demonstrated increased plant growth when fertilized with human urine.[7]

Methods

Creating Solid Synthetic Urine:

In a well-ventilated fume hood with constant airflow, two liters of distilled water were added to a 2.5-liter non-reactive stainless-steel pan containing individual compounds of medical grade synthetic urine as described in *A New Artificial Urine Protocol to Better Imitate Human Urine* (see Table 1).[8] The fume hood was used to provide a constant airflow to enhance the evaporation of the final product and mitigate the risk posed by any volatile compounds created during the process. The pan was placed on a hot plate with a stir bar set at 95 RPM, maintaining the synthetic urine temperature at 40°C. Throughout the process, safety goggles were worn to protect the eyes. The hot plate temperature was adjusted to 70°C to keep the solution at 40°C. One component, uric acid, is unstable in aqueous solution at 40°C. However, this did not affect the overall plant growth. The main nitrogen source, urea, does not decompose at temperatures less than 60°C. A thermometer monitored the temperature until the urine became too viscous to stir (24 hours). Improved control of the temperature could be implemented for future studies, such as the use of a water bath. The stir bar was removed, and the hot plate temperature was set

DOI: 10.36838/v7i8.12

to 50°C for complete liquid evaporation in the fume hood. The evaporation process took an additional 48 hours. Once the urine solidified, it was removed from the pan and ground into a fine powder using a mortar and pestle (see Figure 1 for the final product). CAUTION: Safety goggles and gloves were worn as precautions.

Table 1: Components of Synthetic Urine as described in *A New Artificial Urine Protocol to Better Imitate Human Urine*.[8] The recipe lists essential electrolytes, organic waste products such as urea and creatinine, and various salts that contribute to the physiological and chemical properties of urine.

Compounds	Quantity (g) for 100 ml/ of water CT-AU
Urea CH_4N_2O	1.2012
Uric Acid $C_5H_4N_4O_3$	0.0168
Creatinine $C_4H_7N_3O$	0.0452
Trisodium citrate dihydrate $Na_3C_6H_5O_7.2H_20$	0.1471
Salt NaCl	0.3156
Potassium Chloride KCl	0.2237
Ammonium Chloride NH_4Cl	0.0802
Calcium Chloride $CaCl_2$	0.0333
Magnesium Sulfate Heptahydrate	0.0493
$MgSO_4.7H_2O$	
Sodium Bicarbonate $NaHCO_3$	0.0168
Potassium Oxalate $K_2C_2O_4$	0.0018
Sodium Sulfate Na_2SO_4	0.1278
Sodium Phosphate Monobasic Dihydrate $NaH_2PO_4.2H_2O$	0.0562
di-Sodium hydrogen phosphate dihydrate $Na_2HPO_4.2H_2O$	0.0071

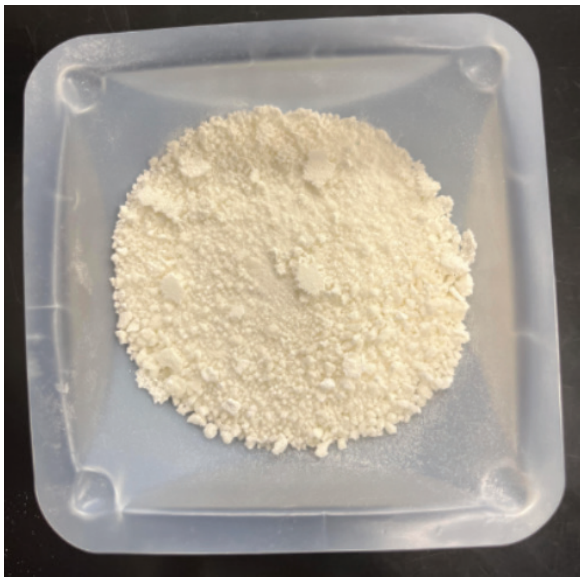


Figure 1: Synthetic urine after dehydration. The dehydration process spanned over 72 hours, resulting in a powdery substance after being ground with a mortar and pestle.

pH of Synthetic Urine:

The pH of the urine solution prior to dehydration was measured and found to be 6.10, within the ideal range for garden cress. pH testing was measured with a calibrated Vernier pH probe. Because varying plants thrive under different pH conditions, additional feasibility titration testing was successfully undertaken using a 100 mL burette with 0.2 M NaOH and HCl. The process was straightforward, and the pH could be adjusted in 0.1 increments up or down. CAUTION: Safety goggles and gloves were worn as precautions.

Serial Dilutions:

Serial dilutions were performed using a 250 mL volumetric flask, distilled water, and a graduated pipette, creating concentrations of 0.025g/mL (equivalent to normal human urine concentration), 0.0025g/mL, and 0.00025g/mL. CAUTION: Safety goggles and gloves were worn as precautions.

Determining Optimal Concentration of Urine:

The individual dilutions listed above were evaluated to determine the optimal urine concentration as a fertilizer. 75g of nitrogen-poor soil was placed into 8 2x2 planters. Then, 25g of nitrogen-poor soil was mixed with one-half teaspoon (2.5 mL) of cress seeds. This mixture was put into all the planters on top of the 75g of nitrogen-poor soil. Then, each pair was fertilized with different fertilizer dilutions. Two planters were watered with 25 mL of water (control), 25 mL of 0.025 g/mL, 25 mL of 0.0025 g/mL, and 0.00025 g/mL. All planters were given equal sunlight and were kept in a stable environment (50% relative humidity and a temperature of 75°F). The planters were watered daily with 30 mL of water. On day 7 of the experiment, the cress was harvested. 0.025g/mL failed to germinate due to the high salt levels. 0.0025g/mL had the highest yield in mass (See Table 2 for results).

Table 2: Masses of cress for determining optimal synthetic urine concentration. The lack of growth in the 0.025 g/mL concentration demonstrates that it would not be feasible to apply pure human urine. Dilution is required.

Control	0.025 g/mL	0.0025g /mL	0.00025g /mL
1.75g	No Growth	3.14g	2.57g

Further Testing with 0.0025g/mL:

Additional tests were conducted to further test the potential of the 0.0025g/mL dilution. 75 grams of nitrogen-poor soil were placed into a 2x2-inch planter. Then, a mixture of half a teaspoon (2.5 mL) of cress seeds with 25 grams of the same soil type was spread over the base layer in each planter. This procedure was replicated across 24 planters. Half of these planters were fertilized with 25 mL of a 0.0025 g/mL synthetic urine solution, the optimal urine concentration. The remaining planters were watered with 25 mL of water to serve as control groups. All planters were placed in a greenhouse under the same conditions as the previous trial (75°F, 50% relative humidity, and equal sunlight). The planters were watered daily with 30 mL of water. On the seventh day of the 14-day growth period, the experimental group of planters was refertilized with another 25 mL of the synthetic urine solution.

The control group planters were again watered with 25 mL of water. Figure 2 shows a notable difference in height between fertilized cress (left) and unfertilized cress (right).

Figure 2: Cress fertilized with synthetic urine (left) and the control group (right). The picture was taken on day 14 of the experiment. Cress fertilized with a 0.0025g/mL concentration demonstrates a fuller, healthier yield compared to the control.

Harvesting Cress:

Harvesting occurred on day 14 using the method described in *Yield and quality of garden cress affected by different nitrogen sources and growing periods*.[9] The height of each planter was measured in cm, cress stalks were cut close to the soil surface, and leftover soil attached to the stalks was removed carefully. The mass of the cress from each planter was measured individually. The yields of each planter were aggregated and exhibited a stark difference in biomass (see Figure 3).

Figure 3: Yield of fertilized cress (left) and unfertilized (right). The large difference in biomass highlights the enhanced growth and higher output of fertilized cress compared to the unfertilized counterpart.

Testing for Bacterial Growth:

Agar plates were created and streaked with *E. coli* K-12 culture to test bacterial growth. Synthetic urine was placed on the plates with a 15µL scoop, spaced evenly with five samples per plate. The plates were placed upside down in an incubator for 24 hours, examined for inhibition zones using a dissecting microscope, and then placed in the incubator for 24 hours. After re-examination, the plates were discarded in an autoclave after sterilization. CAUTION: Safety goggles and gloves were used while handling the Petri dishes to prevent additional bacterial contamination.

Results

Statistical Analyses:

A two-means t-test was conducted, and the results returned as statistically significant. The p-value is less than 0.0001, allowing us to reject the null hypothesis (there is no difference in growth between plants fertilized with the synthetic urine and unfertilized plants) and instead supports the alternative hypothesis (plants fertilized with the synthetic fertilizer have greater growth than unfertilized plants). The city of Harrisonburg, Virginia, has a population of 53,000 people. Assuming that the average person produces 1.4 L of urine daily, enough urine can be collected to fertilize 2,589 acres of cress using the tested regimen after one year.

After conducting a two-means confidence interval test, there is a 95% confidence that the interval 1.96g - 3.99g captures the true average difference between the means of fertilized cress and unfertilized cress. This suggests that, on average, cress fertilized with synthetic urine will be at least 1.96g greater than unfertilized cress.

Two Means Confidence Interval Test:

$$(5.34 - 2.36) \pm 1.8\sqrt{\frac{(1.86)^2}{12} + \frac{(0.6)^2}{12}} = (1.96, 3.99)$$

(difference in means) ± margin of error

As seen in Figure 4 the average mass per planter of fertilized cress is significantly greater than the average mass of unfertilized cress, exemplifying urine's potential as a fertilizer.

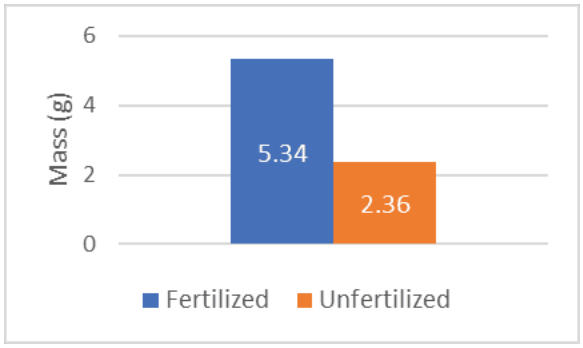


Figure 4: Comparing the average masses per container between fertilized and unfertilized cress. The fertilized cress had an approximately 226% greater yield than unfertilized cress.

Garden Cress:

Lepidium sativum was chosen because of its history in botanical sciences. It is a fast-growing herb that is easily obtainable. Also, as a plant growth model, it is known to be growth-responsive to varying nitrogen levels,[10] as seen in the results. It is a common garnish in soups and sandwiches and adds peppery seasoning. Cress grows best at a pH of 6.0-6.7.[11] After measuring with Vernier pH probes, the synthetic urine met these conditions.

Soil:

Purposefully, the soil used in this experiment was depleted in nitrogen content. It had a history of planting cycles without the addition of nitrogen. Also, it was exposed to the elements, such as rain, for over a decade. Soils are not known to sequester nitrogen, which is a primary reason farmers must resort to the reapplication of fertilizer.[12] Since nitrogen is known for depleting quickly, the soil was tested for nitrogen semi-quantitatively. The nitrogen levels were between N1/Deficient and N2/Adequate. The nitrogen levels rose after applying synthetic urine fertilizer (see Figure 5).

DOI: 10.36838/v7i8.12

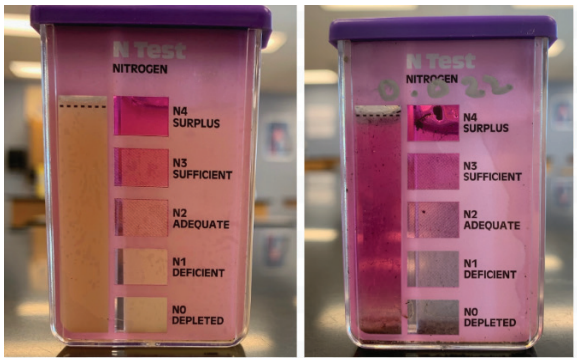


Figure 5: Soil nitrogen levels before and after fertilizer application (deficient/adequate nitrogen levels (left), sufficient/surplus nitrogen levels (right)). The pink color is stronger in intensity in the fertilized soil sample (right), indicating higher levels of nitrogen.

The soil was sent to Virginia Tech, where it was tested at the soil testing lab; the other soil nutrient levels were relatively sufficient. Additionally, the pH was 7.1, near the optimal range of cress growth. See Table 3 for specifics.

Table 3: Properties of the soil used. lb./A = pounds per acre, H = High, VH = Very High, SUFF = sufficient, L = Low, meq = milliequivalent.

Analysis	Result	Rating
P (lb./A)	68	H
K (lb./A)	198	H-
Ca (lb./A)	3680	VH
Mg (lb./A)	243	VH
Zn (ppm)	3.6	SUFF
Mn (ppm)	34.7	SUFF
Cu (ppm)	1.1	SUFF
Fe (ppm)	9.2	SUFF
B (ppm)	0.7	SUFF
S.Salts (ppm)	128	L
Soil pH	7.1	-
Buffer Index	6.60	-
Est.-CEC (meq/100 g)	10.4	-
Acidity (%)	0.0	-
Base Sat. (%)	100.0	-
Ca Sat.	88.0	-
Mg Sat. (%)	9.6	-
K Sat. (%)	2.4	-
Organic Matter (%)	4.1	-

Why use Solid Urine?:

Currently, scientists are studying larger-scale methods that use urine in liquid form, which adds complexity to storage and transportation. Such methods are being analyzed in Sweden and Burkina Faso.[13] However, dehydrating laboratory-grade, synthetic urine leads to a stable shelf substance that is easily reconstituted for use. The dehydrated urine was tested for hygroscopicity and did not gain mass/absorb water over eight weeks (Placed in a 30% humidity environment at room temperature). As a solid, it is easier to transport and store, and does not require to be maintained in a low-humidity environment. This was an interesting finding and somewhat counterintuitive based on the salt content of urine.[14]

Additionally, solid urine's susceptibility to bacterial growth was tested. When diluted in liquid form, bacteria grew unrestrained. However, when solid, the synthetic urine demonstrated a zone of inhibition against *E. coli* K-12 (see Figure 6). This is most likely due to the hypertonic state of the fertilizer, which correlates with the common use of salts as a preservative.

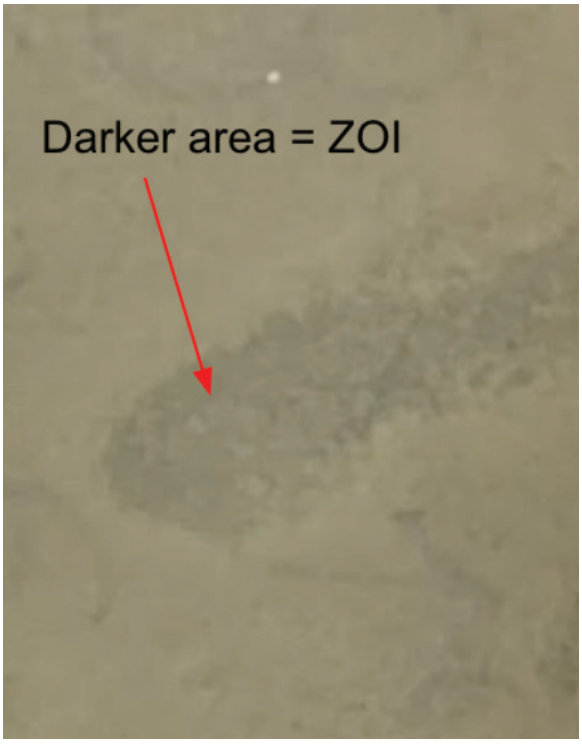


Figure 6: Zone of inhibition around solid urine fertilizer surrounded by extensive *E. coli* K-12 growth. The zone of inhibition extended away from the solid urine location, potentially displaying antibacterial properties.

Other current research emphasizes removing salts to isolate urea. The extraction methods are impractical on a large scale. Multiple steps requiring ethanol evaporation, recrystallization, and vacuum filtration require many resources.[15] This research demonstrates that these are unnecessary steps. Diluting the urine leads to an optimal urine concentration per milligram of water, with enough nutrients to boost plant growth and a low enough salt concentration for plants to germinate and grow.

Synthetic medical-grade urine was used in this study.[8] Common components used to model urine as a fertilizer have not been updated in decades. Currently, scientists continue to use

models that stem from the source: *The Primary Cause of Infection-Induced Urinary Stones*.[16,17] However, the more biologically accurate, medical-grade urine model better helps determine the exact concentration of solid urine to use, resulting in an NPK ratio of the medical-grade synthetic urine using molar calculations of 24:2:5.

Discussion

Future Growth and Implementation:

In nature, a ubiquitous enzyme called urease converts urea into ammonia and carbon dioxide. Because ammonia is a gas under standard temperature and pressure, it can be lost to the environment. Besides the urease produced by garden cress, the model does not include environmental urease. Further studies could add urease to the synthetic urine to simulate a real-world situation. In this setting, it would be possible to identify methods to denature the enzyme to halt degradation before application. Some methods currently in use include acidifying the urine during collection. This approach has been researched in Sweden.[5]

Compared to commercial fertilizer, urine has a relatively low phosphorus content (24:2:5 NPK ratio). Future studies could plan to adjust the phosphorus content to optimize root growth. Additionally, comparing the urine fertilizer to standard commercial fertilizer and studying a more common commercial crop, such as soybeans, would provide greater information on its real-world application. Like cress, soybeans create their urease and would represent a good model, albeit slower growing.

Researchers in Sweden have discovered that a local community produces enough urine to offset its need for commercial fertilizers.[5] However, applying liquid urine to crops is illegal in many E.U. countries because of its odor and promotion of bacterial growth in storage. Additionally, multiple different methods are used to harness waste. Some siphon individual elements from sewage plants, whereas others discard urine entirely, even though "it contains 80 – 90% of the nitrogen (N), 90% of the potassium (K) and 50% of the phosphorus (P) in household sewage."[18] Such problems, however, could be resolved by dehydrating urine. It is odorless, bacteriostatic, and contains an abundance of nutrients. By spreading awareness of the potential of dehydrated urine, harnessing urine as a fertilizer could become a normalized method of fertilization.

Unexpected Outcomes:

Originally, the initial plans were to fertilize the cress with synthetic urine in its solid, unpowdered form. ¼ tsp. and ½ tsp. of urine crystals were placed on top of the soil, and after multiple days, the cress seeds failed to germinate. By this point, the control group was halfway through its growth cycle. After further research, it was determined that it was due to the urine's high salt content.[19] To combat this, further research resulted in a new method of fertilization with serial dilutions. Realistically, inoculating fields with urine's dry form would not be practical as well, as it would be impossible to control its concentration in rainfall-dependent systems. By reconstituting and diluting synthetic urine, the salt concentrations were lowered while maintaining an adequate level of nutrients. In situations where there is poor rainfall, there is a theoretical risk of salt accumulation affecting plant growth. Despite this, Vermont farmers have used urine as a fertilizer for twelve years without noted complications.[20]

Conclusion

This study explores the potential of using human urine as a sustainable nitrogen source for agriculture, focusing on developing a synthetic solid urine model to fertilize *Lepidium sativum* (garden cress). Given the environmental and safety concerns associated with traditional ammonia-based fertilizers produced by the Haber-Bosch Process, this research aims to find a safer, more sustainable alternative. It outlines the creation of a solid urine fertilizer, storage, and testing methods. It highlights the benefits of solid over liquid urine, including reduced risk of microbial growth and easier transportation. The results indicate that garden cress fertilized with the optimal concentration of synthetic solid urine showed significantly higher yields than the control group, proving the effectiveness of urine as a nitrogen source. This approach addresses the challenges of using liquid urine and demonstrates a sustainable path forward in agriculture. With the global population growing and the environmental impact of traditional fertilization methods becoming increasingly unsustainable, this research illustrates the potential of human urine as a fertilizer, demonstrating its potential to contribute to sustainable agriculture practices.

Acknowledgments

We would like to thank Dr. Bill for his guidance and Harrisonburg High School for providing us with laboratory space and materials.

References

1. Capdevila-Cortada, M. Electrifying the Haber–Bosch. *Nature Catalysis* 2019, *2* (12), 1055–1055. https://doi.org/10.1038/s41929-019-0414-4.
2. Jenkins, J. P. Oklahoma City Bombing | Facts, Motive, Timothy McVeigh, & Memorial. *Encyclopædia Britannica*; 2019.
3. El Sayed, M. J. Beirut Ammonium Nitrate Explosion: A Man-Made Disaster in Times of COVID-19 Pandemic. *Disaster Medicine and Public Health Preparedness* 2020, *16* (3), 1–18. https://doi.org/10.1017/dmp.2020.451.
4. Richman, T.; Arnold, E.; Williams, A. J. Curation of a List of Chemicals in Biosolids from EPA National Sewage Sludge Surveys & Biennial Review Reports. *Scientific Data* 2022, *9* (1), 180. https://doi.org/10.1038/s41597-022-01267-9.
5. West, S. INNOVATIONS from SCANDINAVIA: INCREASING the POTENTIAL for REUSE; 2001. https://www.waterfund.go.ke/watersource/Downloads/006.%20Innovations%20Scandinavia.pdf.
6. Yang, D.; Leng, B.; Li, T.; Li, M. Energy Saving Research on Multi-Effect Evaporation Crystallization Process of Bittern Based on MVR and TVR Heat Pump Technology. American Journal of Chemical Engineering 2020, 8 (3), 54. https://doi.org/10.11648/j.ajche.20200803.11.
7. Rajani, V.; Alaka, R.; Rajan, S. HUMAN URINE as a FERTILIZER-A COMPARATIVE STUDY USING SOLANUM

DOI: 10.36838/v7i8.12

LYCOPERISCUM and CAPSICUM SP. https://www.mutagens.co.in/jgb/vol.04/2/14.pdf.
8. Sarigul, N.; Korkmaz, F.; Kurultak, İ. A New Artificial Urine Protocol to Better Imitate Human Urine. *Scientific Reports* 2019, *9* (1). https://doi.org/10.1038/s41598-019-56693-4.
9. Tuncay, Ö.; Eşiyok, D.; Yagmur, B.; Okur, B. Yield and quality of garden cress affected by different nitrogen sources and growing period. *ResearchGate*. https://www.researchgate.net/publication/228346691_Yield_and_quality_of_garden_cress_affected_by_different_nitrogen_sources_and_growing_period.
10. İnne, A.; Kul, R.; Ekinci, M.; Turan, M.; Yildirim, E. Azot Gübrelemesinin Terede (Lepidium Sativum L.) Büyüme, Verim, Nitrat ve Mineral İçeriğine Etkileri. *Yüzüncü Yıl Üniversitesi Tarım Bilimleri Dergisi* 2021, *31* (1), 89–97. https://doi.org/10.29133/yyutbd.823959.
11. Jibrin, D. M.; Namakka, A.; Ibrahim, D.A. Response of garden cress (Lepidium sativum L.) to sowing methods, irrigation interval, and fertilizer rates in Northern Guinea Savannah of Nigeria | *Journal of Agriculture and Environment*. https://www.ajol.info/index.php/jagrenv/article/view/235866
12. Delgado, J. A. Quantifying the Loss Mechanisms of Nitrogen. *Journal of Soil and Water Conservation* 2002, *57* (6), 389–398.
13. Gensch, R.; Dorothee, S.; Urine Fertilisation (Large-scale) | *SSWM*. https://sswm.info/water-nutrient-cycle/reuse-and-recharge/hardwares/reuse-urine-and-faeces-agriculture/urine-fertilisation-%28large-scale%29
14. Guo, L.; Gu, W.; Peng, C.; Wang, W.; Li, Y. J.; Zong, T.; Tang, Y.; Wu, Z.; Lin, Q.; Ge, M.; Zhang, G.; Hu, M.; Bi, X.; Wang, X.; Tang, M. A Comprehensive Study of Hygroscopic Properties of Calcium- and Magnesium-Containing Salts: Implication for Hygroscopicity of Mineral Dust and Sea Salt Aerosols. *Atmospheric Chemistry and Physics* 2019, *19* (4), 2115–2133. https://doi.org/10.5194/acp-19-2115-2019.
15. Marepula, H.; Courtney, C. E.; Randall, D. G. Urea Recovery from Stabilized Urine Using a Novel Ethanol Evaporation and Recrystallization Process. *Chemical Engineering Journal Advances* 2021, *8*, 100174. https://doi.org/10.1016/j.ceja.2021.100174.
16. Griffith, D. P.; Musher, D. M.; Itin, C. Urease. The Primary Cause of Infection-Induced Urinary Stones. *Investigative Urology* 1976, *13* (5), 346–350.
17. Dominika Szczerbiec; Katarzyna Bednarska-Szczepaniak; Agnieszka Torzewska. Antibacterial Properties and Urease Suppression Ability of Lactobacillus Inhibit the Development of Infectious Urinary Stones Caused by Proteus Mirabilis. *Scientific Reports* 2024, *14* (1). https://doi.org/10.1038/s41598-024-51323-0.
18. Johanson, M.; Lennartsson, M. *Sustainable Wastewater Treatment for Single Family Homes.* ICR. https://www.ircwash.org/sites/default/files/320-99SU-17936.pdf.
19. Court, M. N.; Stephen, R. C.; Waid, J. S. Toxic Effect of Urea on Plants: Nitrite Toxicity Arising from the Use of Urea as a Fertilizer. *Nature* 1962, *194* (4835), 1263–1265. https://doi.org/10.1038/1941263a0.
20. Warner, B. Why Vermont farmers are using urine on their crops. Bbc.com. https://www.bbc.com/future/article/20250227-the-vermont-farmers-using-urine-to-grow-their-crops.

Authors

Thomas Bill is a junior at Rocktown High School and Massanutten Regional Governor's School for Environmental Science. He is interested in Biochemistry and Environmental Science, which are his intended majors for college.

Leslie Yang is a junior at Rocktown High School and Massanutten Regional Governor's School for Environmental Science. She plans on pursuing microbiology after graduation.

IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

A Study on the Effects of Taekwondo on the Physical and Mental Health of Teenage Trainees

Eric Ju

East Chapel Hill High School, 500 Weaver Dairy Rd, Chapel Hill, NC, 27514; ericwyverns@gmail.com
Mentor: Hyunjun Ju

ABSTRACT: The purpose of this study is to compare the training goals and methods of Taekwondo studios in the United States and Korea, and to identify the effects of Taekwondo on the physical and mental health of teenage trainees. Taekwondo, as a traditional Korean martial art, emphasizes not only physical training but also mental discipline. As Taekwondo has expanded globally beyond Korea, comparing the training methods of the birthplace of Taekwondo (Korea) and the United States can help identify ways to improve the health of trainees. For this study, surveys were conducted with 60 high school Taekwondo trainees in the U.S. and Korea, and interviews were also held with 6 Taekwondo instructors and 12 trainees. The results showed that the training objectives, content, duration, and methods differ between the two countries. In the U.S., Taekwondo focuses primarily on physical training as a sport, whereas in Korea, it integrates traditional cultural values, emphasizing not only physical training but also mental aspects such as ethics, filial piety, and order. The findings of this study suggest ways to incorporate Korea's approach to mental health training into the U.S. Taekwondo practice, offering methods for balanced physical and mental development in adolescents.

KEYWORDS: Taekwondo, Physical Health, Mental Health, Stress Management, Self-Esteem.

Introduction

Taekwondo Growing as a Global Sport:

Taekwondo is a sport practiced by more than 100 million people worldwide and is actively practiced in 213 countries.[1] Especially, in the United States, 30,000 instructors are running Taekwondo academies, and the cumulative number of practitioners has exceeded 4 million.[2] Taekwondo studios in Korea not only focus on teaching Taekwondo and promoting health but also emphasize values such as filial piety, morality, and mental discipline, which contribute to mental health. Furthermore, they foster networking effects between students, parents, and the local community. Taekwondo has also had a positive impact on helping students grow into responsible members of society. The role of Taekwondo has expanded beyond physical training to include the cultivation of virtues such as courtesy, honesty, tolerance, and self-discipline, which are essential for forming social relationships. Additionally, its benefits for both physical and mental health have increased social expectations.[3] The social values of Taekwondo, which contribute to creating responsible members of society and promote balanced physical and mental health, are recognized internationally.[4]

In the United States, Taekwondo has a significant impact on both the community and families. Many areas lack evening programs after school, so for dual-income families, Taekwondo academies are being used as after-school childcare facilities. In response to this need, these academies are continuously developing programs at the level of after-school education centers,[5] while Taekwondo's popularity among adults in the U.S. fosters family bonding, as fathers and sons, as well as grandfathers and grandsons, train together. This not only strengthens family ties but also helps foster the development of proper family relationships, including values such as filial piety, respect, and honor.

The effects of Taekwondo training on physical and mental health:

Taekwondo is known for its combination of physical exercise, mental discipline, and self-defense techniques.[6] Emphasis on high-intensity training and dynamic movement patterns of Taekwondo contributes to numerous physical health benefits. According to the results of a meta-analysis of various studies on the physical effects of Taekwondo, Taekwondo training during adolescence has a positive impact on muscle strength, muscular endurance, cardiovascular endurance, explosive power, agility, and flexibility. Since Taekwondo is a martial art that involves full-body training, it actively stimulates the organs and functions of various respiratory and circulatory systems within the body, aiding in metabolism.[7] In particular, it can significantly strengthen the respiratory system. Additionally, the harmonious functioning of the nervous system is required for the full-body muscle activity and quick, accurate decision-making and execution, which is actively engaged and enhanced, highlighting the unique characteristics of Taekwondo.[8] Additionally, it is effective in improving obesity and stimulating the secretion of growth hormones.[9] There are many studies on the mental effects of Taekwondo, and research has shown that Taekwondo training has a positive impact on reducing stress and enhancing self-esteem in adolescents. It also contributes to improving mental quality of life by alleviating symptoms of depression and reducing anxiety.[10] In this way, Taekwondo helps develop both physical abilities and social skills. Through Taekwondo training, mental development is supported in

DOI: 10.36838/v7i8.13

terms of cognitive aspects and emotional stability, which enhances quality of life and plays a key role in developing social relationships.[11] The effects of Taekwondo are well reflected in the fundamental goals of Taekwondo education. Taekwondo education aims to intentionally cultivate individuals who are mentally, physically, and socially desirable. Each Taekwondo movement or educational activity is assessed for its educational value, and the teaching and learning processes are evaluated to determine if they help transform practitioners into individuals with positive traits. This value-based education is essentially a process of intentionally changing one's character, abilities, attitudes, habits, and beliefs. Ultimately, Taekwondo training during adolescence can contribute to maintaining a healthy state both physically, mentally, and socially.[12]

The Necessity of Research:

According to previous studies, Taekwondo training helps improve both physical and mental health; however, the degree of improvement may vary depending on the training objectives, methods, and country. In South Korea, the birthplace of Taekwondo, emphasis is placed on social maturity, such as courtesy, filial piety, and morality, while Taekwondo studios in the United States focus more on physical health.[13] Certainly, some studies show Taekwondo training in the United States also contributes to mental health. Adolescents who practiced Taekwondo in the U.S. scored higher in stability, autonomy, sociability, activity, and courtesy compared to those who did not train. Additionally, sub-factors of mental health, such as interpersonal sensitivity, depression, hostility, anxiety, and somatization, were found to be lower in those who practiced Taekwondo.[14]

However, the extent to which Taekwondo training improves mental and physical health may vary depending on the goals and programs of Taekwondo education. Comparing the Taekwondo training programs of the two countries highlights the need for further research. Based on an analysis of Taekwondo programs from six studios in South Korea and the United States, Table 1 highlights the differences in program duration and session content. The Taekwondo program in South Korea typically lasts 60 minutes. The first 5-10 minutes are dedicated to light warm-up exercises, followed by Taekwondo training. The last 5-10 minutes involve discipline and communication between the instructor and the trainees before the session ends. In the United States, the classes are usually 45 minutes long, with 5 minutes allocated for warm-up and 40 minutes for training. The difference in mental health-related factors lies in the last 5-10 minutes, which are excluded from Taekwondo training in the U.S. The last session of the Taekwondo program in Korea goes beyond everyday conversation and serves an educational purpose. The content of the last session includes activities such as meditation, moral and etiquette education, gratitude towards parents, and leadership training. To conduct these sessions, instructors often receive specialized training from professional organizations or obtain relevant certifications.

Table 1: A comparison of Taekwondo programs in South Korea and the United States.
South Korean programs are longer and emphasize both physical and mental training, while U.S. programs focus more on physical training.

Program/Day		United States (45 minutes)	South Korea (60 minutes)
Warm-up	Stretching, Joint rotations, Breathing exercises	5 min	5 ~ 10 min
Training	Forms, Attack and Defense, Board breaking, Demonstration, Cool-down exercise	40 min	45 min
Supplementary training	Meditation, Self-reflection, Communications Discipline (Etiquette, Filial piety, Self-esteem)	-	5 ~ 10 min

* Analyzed US (NC, Texas) and Korea (Seoul, Gyeonggi Province) Taekwondo Studios' programs.

* Refer to studies on Taekwondo programs from the United States[15] and South Korea.[16]

Methods

The object of study:

The study was conducted at 6 Taekwondo studios in the United States and South Korea. To ensure the objectivity of the data, the research was carried out across three studios in Texas and North Carolina in the U.S., and three studios in Seoul and Gyeonggi Province in South Korea, with both surveys and interviews being conducted. As shown in Table 2, the surveys and interviews were conducted with high school Taekwondo trainees and their instructors. Among the survey participants of trainees, 36 were male and 14 were female, aiming to include a diverse range of perspectives. However, due to the low number of female high school trainees, achieving gender balance was difficult.

Table 2: Interviewee and survey participant information from South Korea and the United States.
Participants from both countries showed a similar demographic balance, but Korean trainees had a longer average training period (5.2 years) compared to those in the U.S. (3.4 years).

	Country	Trainees				Instructors			
		No	Age	Gender	Training Period	No	Age	Gender	Coaching Period
INTERVIEW	US	1	15	Female	3 years	1	41	Male	11 years
		2	16	Female	2 years				
		3	16	Male	2 years	2	36	Male	7 years
		4	17	Male	4 years				
		5	17	Male	7 years	3	33	Female	3 years
		6	18	Male	5 years				
	Korea	7	15	Female	5 years	4	42	Male	15 years
		8	15	Male	8 years				
		9	16	Female	3 years	5	31	Male	5 years
		10	16	Male	6 years				
		11	16	Male	7 years	6	29	Female	5 years
		12	17	Male	10 years				

SURVEY	Country	Headcount	Male / Female	Average Age	Average Training
	US	30	19 / 11	16.5	3.4 years
	Korea	30	23 / 7	16.0	5.2 years

Research tools:

This study combined qualitative and quantitative research methods to compensate for the lack of survey data. The qualitative research focused on conducting interviews with Taekwondo instructors and trainees in the United States and South Korea to identify the specific outcomes of physical and mental improvements felt by the participants and the purpose of the Taekwondo programs. Interviews were conducted face-to-face whenever possible based on the questionnaire in Table 3, and for Korean students who could not meet in person, phone and email interviews were conducted. For the

U.S. instructors and trainees, questions were sent via email a few days before the interview, which was conducted in person. The survey was distributed via email not only to the interview participants but also to high school trainees from the same Taekwondo studio. The email provided a detailed explanation of the study's purpose and objectives, the definitions of physical and mental health, and the differences between Taekwondo programs in South Korea and the United States. The study proceeded only after obtaining the participants' consent and in compliance with research ethics.

Table 3: Interview Questions for Taekwondo Instructors and Trainees.
The questions explore Taekwondo's health effects and identify areas for improvement.

No	Interview Questions
1	What is the purpose of training Taekwondo?
2	What are the effects of Taekwondo training on physical health?
3	What are the effects of Taekwondo training on mental health?
4	Are there training programs specifically designed to improve physical/mental health?
5	What aspects of Taekwondo should be improved to enhance physical/mental health?

* The survey questionnaire is attached in a separate file.

Data Analysis:

Based on the analysis results of the 5-point scale survey, the differences between Taekwondo programs in the United States and South Korea were identified, and further research results were derived by incorporating the interview findings. The interview results were obtained by integrating the content recorded during face-to-face interviews and the responses received via email, from which key responses for each question were extracted. Discrepancies between the survey results and interview responses were addressed through additional interviews to identify the underlying causes. To explore the relatively under-researched impact on mental health in more depth, survey items related to self-esteem,[17] sense of belonging, and stress relief, which had been frequently highlighted in previous studies,[18] were added. However, variables related to mental disorders, such as depression and anxiety, were excluded as they do not align with the research objectives. Additionally, correlation analysis of the survey results was conducted using SPSS, and the study identified how factors such as country, age, gender, and training duration impact health improvement.

Research Ethics:

This study complied with the following research ethics. First, before conducting interviews and surveys, participants were explained in detail about the purpose of the study, the plan to use the results, and the sharing of the results, and their voluntary consent was obtained. Second, after making sure that the participants understood prior information before the interview, they were guided not to include responses if they felt uncomfortable during the interview or did not want their responses to be used in the study. Third, for the privacy protection of interviewees, the interview was conducted in a personal environment, and all names were anonymized. Finally, it was informed that the interview and survey results would only be used in the study, and it was communicated that they would not be disclosed to the outside for purposes other than the study.

Results

The significance between variables and outcomes was confirmed through correlation analysis, and the research results are organized in sequence based on the interview and survey responses. Additionally, the interviewees' thoughts and opinions on Taekwondo training are directly quoted. The survey questionnaires on mental health utilized questions related to stress, self-efficacy, and psychological stability from the previously validated questionnaires, Youth Mental Health Risk Assessment (YMHRA)[19] and Mental Health Inventory for Adolescents (MHIA).[20] The mental health level was derived from the average scores of these three questionnaires. According to the correlation analysis in Table 4, the effect of Taekwondo on physical health showed no significant differences across country, gender, or training period. On the other hand, the effect on mental health showed a strong correlation with country (0.603) and training period (0.683).

Table 4: Results of the analysis on the correlations between the variables (N=60).
Notable correlations include a moderate positive relationship between mental health and training period (0.683) and between mental health and country (0.603).

	1	2	3	4	5
1. Physical Health	1				
2. Mental Health	0.122	1			
3. Country (US/KOR)	-0.067	0.603	1		
4. Gender (Male/Female)	-0.117	-0.256	-0.145	1	
5. Training Period	0.097	0.683	0.431	-0.445	1

A T-Test was conducted, as shown in Table 5, to confirm the difference in perceptions between the United States and Korea on the effects of Taekwondo on health. Both groups reported similarly high physical health scores (U.S. M = 4.57, Korea M = 4.50), with no significant difference, $t(58) = 0.51$, $p = .61$, suggesting a shared recognition of its physical benefits. However, Korean participants reported significantly higher mental health benefits (M = 4.30) than U.S. participants (M = 3.73), $t(57) = -5.76$, $p < .001$. This may reflect cultural differences in the way Taekwondo is taught and understood. Overall, physical benefits are consistently acknowledged in both countries, but perceptions of mental health vary by context. American Taekwondo programs can benefit from introducing Korean mental health programs.

Table 5: T-Test results comparing health perceptions between the U.S. and Korea (N=60).
The t-test revealed no significant difference in physical health between the U.S. and Korea. However, mental health perceptions were significantly higher in Korea compared to the U.S.

	Physical Health		Mental Health	
	US	KOREA	US	KOREA
Mean	4.57	4.50	3.73	4.30
Variance	0.25	0.26	0.17	0.12
Number of Observations	30	30	30	30
Hypothesized Mean Difference	0.00		0.00	
Degrees of Freedom	58.00		57.00	
t-statistic	0.51		-5.76	
P(T <= t) (One-tailed test)	0.31		0.00	
t Critical (One-tailed test)	1.67		1.67	
P(T <= t) (Two-tailed test)	0.61		0.00	
t Critical (Two-tailed test)	2.00		2.00	

DOI: 10.36838/v7i8.13

The Effect of Taekwondo on Physical Health:

Survey results in Table 6 showed that both American and Korean trainees agreed that Taekwondo training has a high impact on physical health. American trainees rated it 4.6 out of 5, while Korean trainees rated it 4.5, showing little difference. These results are consistent with previous studies that have highlighted the physical health benefits of Taekwondo. In the interviews, instructors from both countries emphasized that improving physical health is a key goal of Taekwondo training, and both the training methods and the training duration are focused on enhancing physical health.

Table 6: Survey results on the effect of Taekwondo on physical health (N=60, US and Korea).
The survey results show Taekwondo similarly improves physical health in the U.S. and Korea.

Survey Question	Average (US / KOR)
Taekwondo training helps in improving physical health.	4.5 (4.6 / 4.5)

* Five-point rating scale: 1point (Strongly Disagree) ~ 5point (Strongly Agree)

"Taekwondo training is systematic and diverse compared to other school sports. The warm-up exercises vary depending on the day's training content, starting with joint-specific preparations and muscle relaxation. When practicing forms, flexibility exercises are emphasized, while during breaking practice, the focus is on wrist and ankle care. Since Taekwondo is a martial art, there is a risk of injury, which is why the training is conducted according to a well-prepared process, guided by the instructor, to ensure safety." (US, Trainee)

"After becoming a high school student, the amount of academic work increased, and the opportunity to exercise decreased. Taekwondo training is the only time during the day when I can work out intensely and sweat. Although it is high-intensity exercise, I can train systematically and safely under the guidance of the instructor, and I don't feel much muscle pain or fatigue afterward." (Korea, Trainee)

"All training follows a manual. To prevent injuries and maximize exercise effectiveness in a short period, we apply years of accumulated experience and training methods. The Taekwondo Association shares these training methods with Taekwondo schools worldwide, and new training techniques are studied through seminars and forums." (Korea, Instructor)

The Effect of Taekwondo on Mental Health:

While both the United States and South Korea showed positive survey responses regarding the impact of Taekwondo on physical health, there was a significant difference in the reported effects on mental health. As shown in the results of Table 7, the United States scored 3.7 out of 5, while South Korea scored 4.3, showing a 0.6-point difference. When comparing the four elements of mental health, there was little difference between countries in terms of stress relief, but significant differences were observed in self-efficacy, sense of belonging, and psychological safety. Based on the interview results, stress relief can be seen as a secondary effect of physical exercise, while the other elements are developed through mental training or extra activities.

Table 7: Survey results on the effect of Taekwondo on mental health (N=60, US and Korea).
The survey results show that Taekwondo training positively impacts mental health, with Korean participants reporting higher ratings, especially in self-esteem and sense of belonging.

Survey Question	Average (US / KOR)
Taekwondo training helps in relieving stress.	4.2 (4.1 / 4.3)
Taekwondo training helps in enhancing self-esteem.	4.0 (3.7 / 4.3)
Taekwondo training helps in improving a sense of belonging.	4.1 (3.6 / 4.5)
Taekwondo training helps in promoting psychological safety.	3.8 (3.5 / 4.1)
Average	4.0 (3.7 / 4.3)

* Five-point rating scale: 1point (Strongly Disagree) ~ 5point (Strongly Agree)

Both in the United States and South Korea, the opinion was that the stress relief effects of Taekwondo training are similar to those of general exercise, such as improved circulation, a sense of refreshment, and the promotion of hormones like dopamine and serotonin.[21]

"There have been many studies showing that exercise has a positive effect on stress relief, and the same effect is seen in Taekwondo. When considering the unique stress-relief benefits of Taekwondo, many believe that the lack of stress from competition, due to the focus on self-discipline, contributes to this. Additionally, stress is relieved through safely conducted activities like breaking and striking. Exercising, sweating in a healthy way, and completing the targeted workout all contribute to a high level of stress relief." (US, Instructor)

Self-esteem, sense of belonging, and psychological safety are significantly higher in South Korea compared to the United States, with many attributing this difference to the additional mental training conducted alongside physical training. In Korean Taekwondo schools, mental training typically takes place on specific days of the week, focusing on themes such as confidence, leadership, and ethical awareness, and lasts about 10 minutes each day. In addition, instructors and trainees, as well as fellow trainees, share their daily experiences and discuss and consult on psychological difficulties together. Specialized programs like meditation are also considered helpful for mental health.

"After the workout, a lot of preparation is made for the last 10 minutes of mental training. Especially since high school trainees are in their adolescent years, conversations and advice need to be handled carefully. Nevertheless, most trainees listen attentively and empathize. Rather than one-sided lectures, positive examples and desirable directions are shared naturally. Sometimes, even if it's unrelated to the theme, practitioners share challenges they faced during the day, and the group helps each other come up with solutions." (Korea, Instructor)

"The meditation conducted at the end of Taekwondo training is the best part. Listening to calming music while the instructor shares thoughts to wrap up the day helps me feel at peace. I've heard that the instructors have obtained certifications in meditation. Even everyday conversations bring a sense of calm to the mind. Listening to my friends' struggles and talking with them helps us comfort and support each other. Sometimes, conversations with the instructor and friends feel more comforting than with my parents." (Korea, Trainee)

Thematic analysis of interview data from 18 Taekwondo instructors and trainees identified four recurring themes: structured and safe training, physical stress relief, mental stability

and psychological safety, and a sense of community. Based on the results, Korean respondents emphasized mental training and emotional support integrated into Taekwondo training, whereas American respondents emphasized the physical benefits of regular physical activity. These differences stem from cultural and program contexts, suggesting that personalized training programs and policy designs are needed to achieve maximum benefits in each setting.

Table 8: Thematic Coding of Interview Responses on Taekwondo's Health Effects (N = 18).
The coding results show that Korean respondents emphasized both physical and mental training, while the US respondents focused on physical benefits.

Theme	Similarities and Differences	Keywords
1. Structured and safe training	Common theme (emphasized in both Korea and the US)	Systematic training, safety, instructor guidance
2. Physical stress relief	Common theme (exercise-related stress relief emphasized by both)	Sweating, healthy exercise, stress reduction
3. Mental stability and psychological safety	More structured and emphasized in Korea	Meditation, mental training, emotional care
4. Sense of community	Strongly emphasized in Korea	Peer support, empathy, sharing experiences

Discussion

How to improve physical health in adolescents through Taekwondo:

According to a survey conducted on the areas adolescents specifically want to improve among the physical benefits of Taekwondo training - Endurance, Flexibility, Cardiopulmonary function, and Strength which identified in existing research[22] - male practitioners wanted to enhance Strength, with 67% expressing this preference, while female practitioners wanted to enhance Flexibility, with 72% expressing this preference. According to Table 9, there was no significant difference in the demand for physical health improvement between trainees in Korea and the United States. Additionally, 15% of respondents indicated that there is nothing to improve in Taekwondo for physical health, suggesting that some trainees are satisfied with their current training methods.

Table 9: Survey results on the most needed physical training improvements (N=60, US and Korea).
The results show that males mostly need strength training, while females prioritize flexibility.

	Endurance	Flexibility	Cardiopulmonary	Strength	None needed
Male	1 (2%)	3 (7%)	2 (5%)	28 (67%)	8 (19%)
Female	0 (0%)	13 (72%)	1 (6%)	3 (17%)	1 (6%)
Total	1 (2%)	16 (27%)	3 (5%)	31 (52%)	9 (15%)

Since the survey was conducted on adolescents, there was a higher interest in improving appearance and body shape rather than in intense exercises aimed at strengthening basic physical fitness. Because the time available for exercise during the day is limited, there is a strong demand for increasing strength and flexibility through Taekwondo training. When asked about the exercise they would most like to do besides Taekwondo, males selected bodybuilding, and females chose yoga and Pilates, which aligns closely with the survey responses.

"Many adolescent trainees have expressed a desire to strengthen their muscles, so recently, exercises such as push-ups, squats, lunges, and planks have been incorporated into Taekwondo training for about 5 to 10 minutes. To prevent injuries and ensure effective training, all instructors have learned teaching methods from a professional trainer, and the trainees' responses to the strength training have been generally positive." (Korea, Instructor)

"Adding other exercise methods within the limited time of about 40 minutes is not easy. We tend to focus on the various content that needs to be taught within Taekwondo itself. I understand the high demand among adolescents for strength and flexibility training. It seems beneficial to try new exercises that trainees want to incorporate during breaks in Taekwondo training." (US, Instructor)

"The Taekwondo program is well-designed for physical health. Especially, learning martial arts not only strengthens the body but also enhances mental toughness. However, Taekwondo training has a lower intensity compared to outdoor sports like soccer or athletics. Because of this, I considered switching to bodybuilding or personal training for strength training. I've heard that Taekwondo gyms in Korea offer additional strength training programs. It would be great if a strength training program could be added and supplemented at the Taekwondo instructor level." (US, Male Trainee)

"Honestly, Taekwondo is a type of martial art that is male-dominated, and the number of female trainees tends to decrease after high school. While flexibility is required for kicks and movements, and it's not difficult for women to learn, training alongside men can be challenging due to the program's structure. There are many aspects in which Taekwondo needs to change to better cater to women's physical training. If possible, it would be great to create a separate Taekwondo program for women. If the program includes patterns and movements more suitable for women, as well as incorporates yoga or Pilates exercises, which are highly requested by women, I believe more female trainees would participate." (Korea, Female Trainee)

How to improve mental health in adolescents through Taekwondo:

Some studies suggest the physical training of Taekwondo helps reduce stress and alleviate depression, improving mental health,[23] but there are not many existing studies on the effects of Taekwondo's mental training programs. In this study, through interviews with Korean instructors, mental health-related content was identified from the supplementary programs currently offered in Taekwondo studios, and a survey was conducted based on these programs.

Table 10: The definition of mental health programs conducted during Taekwondo training in Korea.
Definitions were provided to help survey respondents understand each program.

	Definition
Meditation	Focusing the mind in a calm state to restore mental stability and comfort.
Leadership	Sharing examples of good leaders and discussing the virtues should learn.
Social skills	Lectures on gratitude towards parents and teachers, and on ethics and morals.
Relationships	Sharing the daily routine and having an open conversation to solve concerns.

The survey results showed no significant differences based on gender or country. For trainees in the United States, where mental health improvement programs were not available, brief descriptions of each program were included in the survey, as shown in Table 10.

DOI: 10.36838/v7i8.13

Table 11: Survey results on the most needed mental training improvements (N=60, US and Korea).
The results show that meditation is the top mental training need, followed by relationship improvements and leadership.

	Meditation	Leadership	Social skills	Relationships	None needed
Male	19 (45%)	7 (17%)	1 (2%)	12 (29%)	3 (7%)
Female	9 (50%)	2 (11%)	0 (0%)	6 (33%)	1 (6%)
Total	28 (47%)	9 (15%)	1 (2%)	18 (30%)	4 (7%)

The survey results in Table 11 showed that meditation was the most preferred program, with 47%, followed by the relationship between instructors and trainees, which was favored by 30%. In addition, the survey results indicated that leadership and social skills scored somewhat lower, with instructors analyzing that this was due to the survey participants being high school students. It was observed that while social skills and leadership are well-received by younger students, the demand for these programs may be lower among more mature high school students.

"Meditation is conducted twice a week, and most trainees, regardless of gender, enjoy it. Korean high school students are mentally exhausted due to studying and preparing for university entrance exams. Because of this, the focus is on improving mental health through meditation rather than strict discipline or guidance. Some trainees have expressed a desire for daily meditation sessions, so the instructors are working to enhance the quality of the meditation program by studying specialized programs and seeking better content." (Korea, Instructor)

"Through this interview, I was able to understand the differences in programs between Korea and the United States, and I was quite surprised to learn that trainees have a high level of interest in mental health improvement programs. While I would like to include mental health improvement programs, it seems difficult to extend the overall program duration. However, we can complement this through online programs or special seminars, similar to what was used during the COVID-19 period. In particular, meditation has gained significant attention in the United States recently, and I feel there is a pressing need to implement it quickly." (US, Instructor)

"Meditation at the end of training is beneficial, but even a short amount of time for trainees to talk about their daily routines under the guidance of the instructor also contributes to mental health. They exchange thoughts on each other's concerns, and instructors often offer solutions based on their own experiences. In fact, conversations between Taekwondo trainees are much more comfortable than those with parents, teachers, or school friends. The reason the relationships at the Taekwondo studio feel more comfortable than those with school friends or teachers is that there is no competition or conflicting interests among the friends there." (Korea, Female Trainee)

"After hearing about the mental health programs at Taekwondo studios in Korea, I thought it would be great to introduce them in the United States as well. While it's common to meet friends from the same school at Taekwondo studios, there is no relationship building outside of the sport. Also, in the U.S., there are no separate Taekwondo classes for high school students, so it might be difficult to have meaningful conversations based on shared experiences. I strongly support the idea of extending training time and creating a separate class for high school students. Training and communicating with peers of a similar age group would be a wonderful experience." (US, Male Trainee)

Conclusion

This study confirmed that Taekwondo training has a positive effect not only on physical health but also on mental health improvement. Through the analysis of differences by country, it was found that Taekwondo studios in the United States are relatively lacking in mental health programs, and both instructors and trainees were supportive of enhancing mental health improvement programs. According to an open-coding analysis of interview data, both Korean and American trainees recognized Taekwondo training as a systematic, safe, and physically healthy activity. The topics commonly mentioned by the trainees from both countries included systematic warm-up, injury prevention, and physical improvement. However, in terms of mental health, the difference was seen. While Korean trainees emphasized more structured mental training, such as self-esteem, belonging, emotional support, meditation, and group reflection, these factors were relatively less mentioned among American trainees. This suggests a culturally embedded approach to mental health in Korean taekwondo training.

Additionally, the specific needs of high school trainees were identified. Physically, they desired strength and flexibility improvement, while mentally, there was a need for meditation and relationship-building programs. However, to implement mental health improvement programs, a comprehensive overhaul of the Taekwondo programs in the United States is necessary. First, training time should be extended by 5 to 10 minutes, and in addition to the current classes for child trainees, separate classes for high school students should be created. Moreover, instructors will need additional training and knowledge to run mental health improvement programs. To support this, instructors should be provided with educational opportunities to enhance their expertise in areas such as meditation and counseling.[24] Furthermore, collaboration between Taekwondo associations and local studios is essential to support the development of instructors' capabilities.[25] Taekwondo may not be able to meet all the needs of high school trainees. However, by referencing successful examples from other countries, continuously seeking areas for improvement, and embracing change, Taekwondo will become a program that is loved across nations and generations.

The current study provides valuable insights into the impact of Taekwondo training on mental health, future studies plan to apply a more experimental approach. Specifically, I would like to have a stronger understanding of the effectiveness of implementing and evaluating mental health training programs through pre- and post-intervention designs within Taekwondo institutions. This approach will provide clear evidence of the effectiveness of mental health programs in Taekwondo and is expected to help with future training programs and policy proposals.

Acknowledgments

I sincerely appreciate the head instructor of VTS Taekwondo Studio in Korea, the instructors in the United States, and the Taekwondo studio friends who took the time to help with the survey. I would also like to extend a special thank you to my mentor, Hyunjun, for guiding me in the analysis of the survey results and interview methods.

References

1. Ha, M.-J.; Kang, H.-S. Study on Scoring Method for Objective Judgment of Quality Competition. Korean Soc. Sport Policy 2024, 22 (4), 65–79.
2. Bozorov, A. The Genesis of Taekwondo Around the World. AJSSHR 2024, 4, 113–118.
3. Lee, D. The Educational Value Exploration and Improvement Plan for Taekwondo Classes. Ph.D. Dissertation, Korea National University of Education, 2007.
4. Utkurovna, N. R.; Sayfiddinovna, T. K. Korean Martial Art and Taekwondo Health. Science Innov. 2024, 3 (Special Issue 34), 460–463.
5. Kelley, R. M. After-School Martial Arts: A History, Perceptions of Academic Advantage, and Effects on Academic Performance. LSU Doctoral Dissertations, 2019, 4939. https://repository.lsu.edu/gradschool_dissertations/4939.
6. Ju, H.; Han, Z.; Gao, Y. A Meta-Analysis of the Effects of Taekwondo on Physical Self-Concept. Percept. Motor Skills 2023, 130(6), 2582–2602. https://doi.org/10.1177/00315125231200277
7. Fikadu, D.; Eshetu (Ph.D), A.; Gemeda, A. The Effects of Taekwondo Training on Cardiorespiratory Fitness and Flexibility Among Adolescent Fitness Trainees. EJSS 2024, 5, 147-157.
8. Chung, P.; Ng, G. Taekwondo Training Improves the Neuromotor Excitability and Reaction of Large and Small Muscles. Phys. Ther. Sport 2012, 13 (3), 163–169. https://doi.org/10.1016/j.ptsp.2011.07.003.
9. Nam, S.-S.; Lim, K. Effects of Taekwondo Training on Physical Fitness Factors in Korean Elementary Students: A Systematic Review and Meta-Analysis. J. Exerc. Nutr. Biochem. 2019, 23 (1), 17–25. https://doi.org/10.20463/jenb.2019.0006.
10. Yang, J. S.; Ko, J. M.; Roh, H. T. Effects of Regular Taekwondo Exercise on Mood Changes in Children from Multicultural Families in South Korea: A Pilot Study. J. Phys. Ther. Sci. 2018, 30 (4), 496–499. https://doi.org/10.1589/jpts.30.496.
11. Song, Y. C. The Effect of Taekwondo Training Satisfaction on Physical Self-Efficacy and School Life Satisfaction. M.S. Thesis, Kyungwon University Graduate School, 2012.
12. Park, S.-U.; Jeon, J.-W.; Ahn, H.; Yang, Y.-K.; So, W.-Y. Big Data Analysis of the Key Attributes Related to Stress and Mental Health in Korean Taekwondo Student Athletes. Sustainability 2022, 14, 477. https://doi.org/10.3390/su14010477.
13. Choi, H. M.; Lee, S. H. The Development and Application of a Narrative Therapy Program to Promote Adolescents' Mental Health. J. Arts Psychother. 2014, 10 (2), 311–335.
14. Lee, K. H. Relationship of Taekwondo Training to Sociability Development and Mental Health among American Elementary Schoolers. J. Korean Alliance Martial Arts 2010, 12 (3), 205–220.
15. Cha, S. H.; Lee, J. B. Characteristics of Taekwondo School Educational Programs and Analysis of Successful Cases in the United States. Korean Soc. Sports Sci. 2006, 15 (4), 703–710.
16. Lee, T.-H.; Kim, D.-M.; Lee, S.-N. A Comparative Analysis of Preference According to Program Taekwondo Gymnasium. Korean J. Sports Sci. 2012, 21 (1), 131–142.
17. Shin, K.-W.; Kim, S.-I.; Hwang, Y.-S. The Effects of Teenage Taekwondo Training on Self-Control, Self-Efficacy, and School Life Satisfaction. Korean J. Sports Sci. 2019, 28 (4), 807–818. https://doi.org/10.35159/kjss.2019.08.28.4.807.
18. Edenfield, T. M.; Blumenthal, J. A. Exercise and Stress Reduction. In The Handbook of Stress Science: Biology, Psychology, and Health; Fisher, S., Ed.; Springer: New York, 2011; pp 301–319.
19. Sampey, L.; Walzl, D.; Wee, C.; Clafferty, R. Improving the Use of the Mental Health Risk Assessment (MHRA). BJPsych Open 2022, 8 (S1), S180–S180. https://doi.org/10.1192/bjo.2022.500.
20. Hann, D. M.; Luby, J. L. Mental Health Inventory for Adolescents (MHIA): A Validated Tool for Measuring Adolescent Mental Health. J. Adolesc. Health 2004, 35 (6), 478–485.
21. Nam, S. N.; Kim, H.; Park, S. J. The Effect of Long-Term Taekwondo Training on Dopamine, Serotonin, and Stress Hormone Levels in Middle-Aged Women. J. Exerc. Sci. 2009, 18 (2), 247–256.
22. Kim, Y. W.; Choi, J. A.; Oh, S. Meta-Analysis Regarding the Effects of Taekwondo Training on Body Composition and Physical Fitness. Kukkiwon Taekwondo Res. 2018, 9 (3), 97–113.
23. Moore, B., Dudley, D., & Woodcock, S. (2020). The effect of martial arts training on mental health outcomes: A systematic review and meta-analysis. Journal of Bodywork and Movement Therapies, 24(4), 402–412. https://doi.org/10.1016/j.jbmt.2020.06.017
24. Kim, H. The Relationship between Job Satisfaction and Professional Satisfaction of Taekwondo Instructors; Master's Thesis, Cheongju University, Chungcheongbuk-do, 2007.
25. Ju, H.; Lee, E. The Improvement of Total Rewards and Working Conditions for Taekwondo Instructors in the COVID-19 Era: A Comparative Study Between the U.S. and Korea. J. Stud. Res. 2021, 10.

Author

Eric Ju is a 4th-degree black belt with 10 years of Taekwondo training in Korea and has served as an assistant Taekwondo coach in the U.S. He is highly interested in biology, biopharmaceuticals, and environmental science, and plans to become an expert in the healthcare sector in the future.

DOI: 10.36838/v7i8.13

IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Can Exercise and Self-care Help Manage Stress and Performance?

Megan Chung[1], Gina Hwang[2]

1)Seoul Foreign School, 39 Yeonhui-ro 22-gil, Seodaemun-gu, 03723, Seoul, Korea; megannchung111@gmail.com
2)UWC Southeast Asia (Dover Campus), 1207 Dover Rd, 139654, Singapore

ABSTRACT: Stress is a global issue that negatively affects individuals and society. Effective stress management is crucial because it can prevent individuals from experiencing negative health conditions such as depression and anxiety, and detrimental workplace behaviors such as suboptimal performance. As physical and mental well-being are crucial facets of a healthy society, this research emphasizes the importance of developing effective stress-management methods to promote individual well-being and performance. Specifically, this study focuses on exercise and self-care practices and how they can reduce stress, which can further help manage individual performance. By examining the police force, a highly stressful occupation, this study found that individuals who exercised and practiced self-care were more likely to have lower stress, which can help maintain their performance. Furthermore, the findings may guide individuals and organizations in promoting a healthy lifestyle and workplace to prevent stress from escalating and maintain high performance levels.

KEYWORDS: Behavioral and Social Sciences, Stress, Exercise, Physical Care, Task Performance.

Introduction

Stress has become a universal experience, affecting millions of people worldwide. According to the World Health Organization,[1] stress-management is crucial to an individual as it can be a risk factor for major physical or mental problems such as heart disease, depression, and/or anxiety disorders.[2] Over 77% of people experience stress, and it has affected their physical and mental health.[2] Especially, high-stress occupations, where individuals face long hours, tight deadlines, and high expectations, have been leading to both physical and psychological health issues globally. As stress continuously takes a toll on health problems across the world, finding effective ways to mitigate the impact has become the main issue today. The constant pressures of everyday life can often spill over to personal life, leading to growing cycles of stress, making it harder to recover to a stable mental state. With the compiled amounts of stress, individuals may find themselves in an emotionally draining state, which can significantly impact their ability to perform effectively in personal and professional settings. Among the various stress reduction strategies available, methods such as exercise and physical care have the potential to alleviate people's stress levels.

Exercise, particularly physical activities such as aerobic exercises and workouts, has been argued to have positive impacts on mental and physical health. These activities can promote relaxation and balance within the human mind and body while maintaining overall well-being. Numerous studies have shown how regular physical activity can reduce depression, anxiety, and stress levels.[3,4] Engaging exercises such as running, swimming, and cycling trigger the body to release endorphins in the brain to act as mood elevators and natural painkillers. These chemical messengers counteract the negative emotions associated with stress and make the body more resilient to stressors in daily life. Additionally, physical activities themselves allow individuals to focus on the activity itself, which functions as a distraction from external pressures or worries that may be adding to stress. Workouts may also lead to a sense of achievement, which fosters greater self-esteem and confidence. Hence, as further research is conducted to emphasize the importance of exercise and in stress-management, it is significant to explore how these practices can be incorporated into daily life routines.

Furthermore, physical care, such as proper nutrition and hydration, can also be integral in reducing stress effectively.[3] Maintaining a healthy diet and planning self-care practices can improve cognitive function and the wellness of others.[3] From this perspective, the current study can help further understand the relationship between stress and physical care and exercise by examining the responses of police officers in South Korea. Using a profession that is known for its high levels of stress and physical demands, they will be able to examine how exercise and physical care can reduce the negative effects of stress. This particular group will provide insight into identifying effective stress-managing strategies as well as recommendations on structuring a healthy workplace for first responders. Through this research, future programs can be developed to maintain stress levels amongst individuals and organizations on how they can approach personal health and professional performance in the long term.

The impact of stress highlights the importance of finding effective stress-management strategies that can promote the well-being of individuals. Surprisingly, due to the lack of empirical studies, the full extent of how physical care and exercise can reduce stress remains underexplored. Subsequently, this research is crucial to explore the interventions that not only address symptoms of stress but also prevent escalating into chronic conditions. The relationship between exercise and

DOI: 10.36838/v7i8.14

physical care with stress is complex, but investigating the relationships of these factors will help enable us to understand the importance of stress management. Therefore, it can be beneficial to help individuals cope with stress while creating a sustainable lifestyle that will further help individuals perform at high levels.

Hypothesis development

Exercise, as a form of physical activity, is a key component of stress management. Aerobic exercises such as running, dancing, swimming, as well as strength training can be effective in stress reduction. The repetitive movements can promote a sense of ease as they increase an individual's heart rate and breathing, which then stimulate mood elevators by shifting the focus away from stressful thoughts. These forms of exercise produce endorphins, which help alleviate mood and provide a sense of relaxation to improve overall cardiovascular health and negative emotions associated with stress.[5]

Exercise contributes to stress relief through stress inoculation by exposing the body to build resilience over time through increased heart rate and muscle tension.[6] During exercise, heart rates increase, which helps the body learn to manage the physiological responses more efficiently.[6] After the exercise is completed, the body goes into a relaxed state, which can then help an individual's ability to manage stressors in life. Moreover, the mind-body connection is significant in coping with stress. The mind and body are interconnected, which impacts how an individual manages stress. When stress affects the mind, it can influence symptoms such as muscle tension, fatigue, and other physical health problems.[4] Thus, we hypothesize the following:

Hypothesis 1: Exercise will be negatively related to stress.

In concert with the stress-reducing properties of exercise, physical care, particularly through constant attention to nutrition and hydration, assumes a significant role in equipping the body to effectively manage stress. Physical care encompasses practices related to nutrition, hydration, and exercise. It is considered a foundation of mindful self-care because it highlights the connection between physical well-being and mental health.[3] A balanced and nutrient-rich diet provides the body with the essential building blocks required for optimal functioning, particularly when confronted with stress. Nutrition, for instance, consists of eating a variety of nutritious foods as well as maintaining sugar levels, nutrient deficits, and inadequate or excessive energy intake.[3] The body's physiological response to stress requires insufficient nutrient intake and poor dietary habits, which can make stress more difficult to manage. Specifically, low iron intake and/or vitamin D levels increase the vulnerability to stress hormones and neurological functions.[3] Similarly, hydration involves drinking the recommended water intake to have healthy functioning in the body. (1.2 L per day) Dehydration can reduce physical and mental health performance due to fatigue, headache, and decreased concentration.[3]

Nutrition and hydration can reduce stress through several factors. Proper nutrition and hydration provide the human body with vitamins and minerals needed to maintain physiological processes and to respond effectively to stressors. The imbalance of nutritious foods, such as high blood sugar, can impair mood regulation and the stability of emotions.[3] Stress hormones, such as cortisol, are directly involved in regulating stress, and neurotransmitters like serotonin can promote feelings of well-being and happiness.[6] Nutritional support can help manage cortisol and serotonin levels through foods with high sources of tryptophan. The maintenance of a balanced nutritional diet and hydration is displayed through the interactions between stress hormones and neurotransmitters. When the body is in a healthy nutritional state, serotonin can be produced, meaning the body may respond and withstand stress more efficiently. Hence, having a balance of nutrition and hydration is optimal for the body's stress-response systems. Hence, we propose the following:

Hypothesis 2: Physical care will be negatively related to stress.

Stress is a complex concept that can be broadly defined as a psychological and physiological reaction to environmental demands that surpass an individual's ability to cope with the situation.[7] Stress is conceptually distinct from strain. Whereas stress pertains to how individuals respond to external demands, strain refers to the consequences resulting from ongoing stress and the coping resources the person possesses to deal with that stress.[8] Stress encompasses a range of cognitive, emotional, and physical responses, such as anxiety, frustration, and fatigue,[9] leading to biological and psychological changes that can harm overall health and well-being.[10] Stress can arise from the interaction between people and their environment, and be further influenced by their cognitive evaluation and coping abilities.[11] Understanding stress can be further refined through the works of Cohen *et al.*,[12] which suggests that stress emerges when an individual's appraisal of environmental demands outweighs their perceived resources to manage those demands. This view emphasizes the subjective nature of stress, which is influenced by individual values, beliefs, and specific circumstances.

Task performance refers to the execution of job-related duties that contribute directly to an organization's goals. This includes behaviors associated with producing goods, delivering services, and supporting management and technical processes.[13] Borman and Motowidlo[14] explain task performance as the set of behaviors that have a direct relationship with the technical core of the organization, coupled with the formal job requirements mentioned in job descriptions. Task performance, while distinct from contextual performance, which Borman and Motowidlo[14] defined as behaviors contributing to an organization's social and psychological environment, enables the organization to achieve its goals. Consequently, task performance is critical to organizational success because it significantly contributes to not only the psychological and social environment of the organization but also its competitiveness and productivity, thus proving essential for achieving the goals of the organization.[15,16]

Previous research has suggested a negative correlation between stress and task performance at the workplace. The Cognitive Activation Theory of Stress (CATS)[11] proposes that stress disrupts cognitive processes essential for goal-directed

behavior, such as memory, problem-solving, and decision-making. For example, high-stress levels decrease concentration levels, leading to frequent mistakes and reduced productivity in task completion. In addition to cognitive responses, stress may also evoke psychological and emotional responses. According to the Transactional Model of Stress and Coping[11] individuals feel stressed when they perceive an imbalance between the demands placed on them and their ability to cope with them. This perception, in turn, triggers an emotional or psychological reaction, such as anxiety, frustration, or helplessness, which further exacerbates the cognitive disruptions caused by stress. From this perspective, Maglio and Campbell[17] highlighted how task performance can be disrupted by stressors that impair focus and efficiency. Stress-induced annoyance and anxiety, for instance, interrupt workflows and reduce the capacity to meet performance expectations. Similarly, Bailey *et al.*[18] found that stress-related anxiety disrupts task execution, particularly in roles requiring sustained focus and multitasking, we propose the following:

Hypothesis 3: Stress will be negatively related to performance.

Methods

Data were collected using two anonymous questionnaires with Likert-scaled items. The first survey (T1) was given in the first week of November 2024, and the second survey (T2) was given in the first week of December 2024. The study sample consisted of police officers, and the response rate was 87% at T1 (261 were returned out of 300 questionnaires sent out), 89% at T2 (231 returned out of 261 questionnaires sent out), and the final sample size was 214, as unusable cases were discarded. Physical care was measured on 5 items based on Hotchkiss & Cook-Cottone,[19] including "I eat a variety of nutritious foods," "I exercise at least 30-60 minutes," "I take part in sports, dance, or other scheduled physical activities," "I practice yoga or another mind-body practice," and a reverse-scored item, "I do sedentary activities instead of exercising." "I drink at least 6 to 8 cups of water" and "I plan my meals and snacks" were excluded as Hotchkiss and Cook-Cottone[19] found them to either have a low factor loading or an overlap with a different item. Exercise was measured using an open-ended item asking, "How many times do you exercise per week?" Stress was measured on 4 items based on Wartig, Forshaw, South, & White[20] and items included "In the last month, I feel I am unable to control the important things in my life," "I do not feel confident I can handle my personal problems," "I feel that things are not going my way," and "I feel difficulties are piling up so high that I cannot overcome them." Moreover, performance was measured on 4 items based on Van Dyne & Lepine[21] and included the items "I fulfill my responsibilities in my job description," "I perform the tasks that are expected as part of my job," "I meet performance expectations," and "I adequately complete my job responsibilities." Besides the frequency of exercise, all of the scales used a 7-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree).

Results

Table 1: The majority of respondents were male, married, and over 31 years old, with most having 6 or more years of tenure. The most common position was inspector, and most respondents held a bachelor's degree.

Variable		%	Variable		%
Gender	Male	78.5	Tenure	5 years or less	29.8
	Female	21.5		6-10 years	25.8
Marital Staus	Unmarried	30.8		11-15 years	13.0
	Married	69.2		16 years or more	31.4
Age	30 and below	16.8	Position	Constable	9.8
	31-40	40.2		Sergeant	21.0
	41-50	29.3		Inspector	29.0
	50 and above	13.7		Lieutenant	27.6
Education	High school graduate	20.1		Superintendent	11.2
	Bachelor's degree	74.7		Chief superintendent	1.4
	Graduate degree	4.2			
	other	0.0			

Table 1 presents the characteristics of the respondents. The majority of respondents were male (78.5%) and had a bachelor's degree (74.7%). About 40% of the respondents were between the ages of 31 and 40, and 29% were between 41 and 50. The tenure of the respondents ranged from 5 years or less (29.8%) to 16 years or more (31.4%). In terms of position, 29% of the respondents were inspectors, and 27.6% were lieutenants. Over two-thirds of the respondents were married (69.2%).

Table 2 illustrates the means, standard deviations, and correlations for the study. As seen in Table 2, exercise and physical care were found to be negatively correlated to stress and positively correlated to task performance. Stress was found to be negatively correlated with task performance.

Table 2: This table presents the means, standard deviations, and correlations of the study variables.

	Mean	S.D.	1	2	3	4	5	6	7	8	9
1. Gender	0.21	0.41	1								
2. Age	39.63	8.69	-0.134	1							
3. Marital status	0.69	0.46	-0.069	0.613**	1						
4. Education	2.57	0.89	0.179**	0.099	0.106	1					
5. Position	3.14	1.21	-0.126	0.761**	0.512**	0.091	1				
6. Tenure	12.56	8.72	-0.125	0.906**	0.543**	0.000	0.817**	1			
7. Exercise	2.32	1.32	-0.155*	-0.021	-0.090	-0.177**	-0.061	-0.006	1		
8. Physical care	3.67	1.35	-0.255**	-0.015	-0.057	-0.122	-0.001	0.033	0.598**	1	
9. Stress	2.86	1.16	0.099	0.033	-0.071	0.167*	0.077	0.026	-0.282**	-0.276**	1
10. Task performance	5.47	0.89	-0.080	0.241**	0.222**	0.019	0.227**	0.237**	0.217**	0.256**	-0.411**

** p < .01.

The hypotheses were tested with SPSS 25 using hierarchical regression analysis. Hypothesis 1 predicted that exercise would have a negative impact on stress. As presented in Table 3, exercise was negatively related to stress ($\beta = -0.148$, $p < 0.05$). Therefore, Hypothesis 1 was supported. Hypothesis 2 proposed that physical care will have a negative effect on stress. As shown in Table 3, physical care was negatively associated with stress ($\beta = -0.143$, $p < 0.05$). Thus, Hypothesis 2 was also supported.

Table 3: The hierarchical regression results showed that exercise and physical care had a significant effect on stress.

Variables	Step 1				Step 2			
	B	Std. Error	Beta	t	B	Std. Error	Beta	t
Gender	0.231	0.196	0.082	1.18	0.049	0.194	0.017	0.251
Age	0.011	0.023	0.079	0.452	0.007	0.023	0.054	0.319
Marital Status	-0.432	0.216	-0.172	-2.003*	-0.499	0.208	-0.199	-2.400*
Education	0.193	0.093	0.148	2.066*	0.154	0.091	0.118	1.698
Position	0.156	0.114	0.163	1.363	0.126	0.110	0.132	1.146
Tenure	-0.010	0.025	-0.075	-0.399	-0.002	0.024	-0.016	-0.089
Exercise					-0.148	0.072	-0.167	-2.038*
Physical care					-0.143	0.071	-0.167	-2.010*
	$R^2 = 0.058$, $F = 2.138$*				$R^2 = 0.141$, $F = 4.191$***			

*p < .05, ***p < .001.

DOI: 10.36838/v7i8.14

Hypothesis 3 predicted that stress would have a negative impact on task performance. In Table 4, stress was found to be negatively related to task performance ($\beta = -0.332$, $p < 0.001$). Therefore, Hypothesis 3 was supported.

Table 4: The hierarchical regression results showed that stress had a significant effect on task performance.

Variables	Step 1				Step 2			
	B	Std. Error	Beta	t	B	Std. Error	Beta	t
Gender	-0.107	0.149	-0.049	-0.715	-0.03	0.135	-0.014	-0.221
Age	0.005	0.018	0.047	0.270	0.008	0.016	0.081	0.518
Marital Status	0.224	0.164	0.116	1.361	0.080	0.150	0.042	0.534
Education	0.005	0.071	0.005	0.069	0.069	0.065	0.069	1.065
Position	0.050	0.087	0.068	0.575	0.102	0.079	0.139	1.291
Tenure	0.007	0.019	0.070	0.377	0.004	0.017	0.038	0.225
Stress					-0.332	0.048	-0.433	-6.960***
	$R^2 = 0.073$, $F = 2.725$*				$R^2 = 0.25$, $F = 9.793$***			

*p < .05, ***p<.001.

We further conducted a bootstrapping test to confirm the mediation effect. The bootstrapping technique was performed with 5000 samples at a 95% confidence interval. The bootstrap results in Table 5 showed that the bootstrapped 95% confidence interval for the indirect effects of exercise on task performance did not contain zero (0.366, 0.1138). Furthermore, the bootstrapped 95% confidence interval for the indirect effect of physical care on task performance did not contain zero (0.342, 0.1083). In sum, stress was found to mediate the relationship between exercise and physical care with task performance.

Table 5: The mediation analysis demonstrated the indirect effect of exercise and physical care on task performance through stress.

Indirect effect on task performance	Effect	BootSE	Bootstrapping Percentile 95 per cent CI	
			Lower	Upper
Exercise	.0708	.0195	.0366	.1138
Physical care	.0681	.0185	.0342	.1083

Discussion

This study aimed to determine the relationships between exercise, physical care, stress, and task performance. Collecting data from police officers in South Korea, the study provides initial insight into how exercise and physical self-care practices can negatively impact stress levels, which then affects individual performance. Given the intense physical demands of police officers, such as exposure to traumatic events and constant public expectations, understanding how physical care and exercise can reduce stress and help police officers not only maintain healthy stress levels and well-being but also perform at high levels.

The data analysis results suggest that exercise and physical care can be quite effective in reducing stress in police officers. The findings of the study align with previous research on how important physical and mental health are in impacting stress levels in humans.[6] Exercise, more specifically aerobic activities, can trigger the release of endorphins, which creates feelings of relaxation throughout the body. This biological response plays a significant role in minimizing the physical effects of stress and helps with preserving homeostasis. Also, physical care practices such as a healthy balance of sleep, nutrition, and hydration can help individuals maintain their stress levels and mental clarity and energy, which allows individuals to perform their duties more effectively while being under pressure. In addition, stress impacts individuals emotionally, which can lead to frustration, anxiety, and depression.[3,19] These emotions can further decrease motivation and affect overall task performance as well. Furthermore, combining exercise and physical care improves job performance and mental health by enhancing cognitive function and boosting mood.

Stress was found to be negatively related to performance. This finding can be explained by the cognitive impact of stress. At the cognitive level, stress interferes with essential functions such as memory, problem-solving, and decision-making. High-stress levels can thus reduce concentration and mental clarity, thereby hindering task efficiency and accuracy. As a result, individuals will be more likely to make mistakes and struggle with productivity when they have higher stress levels. This finding can also be explained by how stress affects individuals emotionally. On an emotional level, stress can elicit feelings such as anxiety, frustration, and helplessness, which further impair task performance. When individuals perceive that the demands placed on them exceed their coping ability, they can become overwhelmed, leading to decreased motivation and efficiency. These emotional responses, in turn, disrupt focus, slow down task completion, making it more difficult to maintain consistent performance levels.

The first potential limitation of this study is that it does not fully account for cultural differences, as the research was conducted in South Korea. Cultural differences may significantly influence perceptions of stress. Prior studies suggest that perceptions of stress, coping mechanisms, and personal control vary across cultures.[22] Therefore, cultural factors may influence the relationship between variables, potentially limiting the applicability of the findings across different cultures among contexts. Future research should explore these relationships within diverse cultural settings to further validate the results. The second limitation is the generalizability of the findings, as the study collected data mainly from male police officers. Perceptions of exercise, physical care, stress levels, and task performance may vary based on gender and occupational background. Further studies should include a more diverse sample, incorporating more females and individuals from various professions to increase the potential for generalization. The third limitation is that exercise was measured using a single item. Therefore, future research employing a multi-item measure capturing frequency, intensity, and type of exercise would enhance construct validity. Finally, individual differences may also play a role in shaping the relationship between the variables. It is well established that differences exist in how individuals deal with stressful encounters,[23] with factors such as personality traits, fitness levels, coping styles, and resilience influencing their responses to stress and engagement in physical activities. Since this study did not fully account for these variations, future research should examine how personal differences can interact with the variables.

Conclusion

In conclusion, the study found that more exercise and self-care practices, such as relaxation and nutritional awareness, can help in reducing stress. Further, as stress can negatively affect

performance, stress management is critical in helping an individual's job performance. Therefore, the study found that stress mediates the relationships between exercise and self-care with performance, which emphasizes the importance of exercise and self-care for individual well-being and job performance.

Acknowledgments

We would like to express our gratitude to Dr. Chung and Dr. Im for their guidance and support in this research. We both took part in the questionnaire development as we searched for the research scales and helped format the questionnaires. Dr. Chung and Dr. Im gave us a tutorial on using SPSS, and after their analysis results, we replicated the results in order to understand how to use the software and to interpret the analysis results.

References

1. World Health Organization, Regional Office for the Eastern Mediterranean. *Stress*; Factsheet, 2023. https://applications.emro.who.int/docs/WHOEMMNH236E-eng.pdf (accessed December 1, 2024).
2. McLean Hospital. Everything You Need to Know About Stress. https://www.mcleanhospital.org/essential/stress (accessed December 3, 2024).
3. Cook-Cottone, C. P.; Guyker, W. M. The Development and Validation of the Mindful Self-Care Scale (MSCS): An Assessment of Practices That Support Positive Embodiment. *Mindfulness*. **2018**, *9* (1), 161–175. https://doi.org/10.1007/s12671-017-0759-1.
4. Davis, M.; Eshelman, E. R.; McKay, M. *The Relaxation and Stress Reduction Workbook*, 6th ed.; New Harbinger Publications, **2008**.
5. Mayo Clinic Staff. Exercise and Stress: Get Moving to Manage Stress. *Mayo Clinic*. https://www.mayoclinic.org/healthy-lifestyle/stress-management/in-depth/exercise-and-stress/art-20044469 (accessed December 10, 2024).
6. Childs, E.; de Wit, H. Regular Exercise Is Associated with Emotional Resilience to Acute Stress in Healthy Adults. *Front. Physiol.* **2014**, *5*, 161.
7. Bourne, L. E., Jr.; Yaroush, R. A. *Stress and Cognition: A Cognitive Psychological Perspective*; IH-045, 2003.
8. Decker, P. J., Borgen, F. H. Dimensions of Work Appraisal: Stress, Strain, Coping, Job satisfaction, and Negative Affectivity. *J. Couns. Psychol.* **1993**, *40* (4), 470-478.
9. Staal, M. A. *Stress, Cognition, and Human Performance: A Literature Review and Conceptual Framework*; National Aeronautics & Space Administration, **2004**.
10. Surachman, A.; Almeida, D. M. Stress and Coping Theory Across the Adult Lifespan. In *Oxford Research Encyclopedia of Psychology* Oxford University Press, **2018**.
11. Biggs, A.; Brough, P.; Drummond, S.: Lazarus and Folkman's Psychological Stress and Coping Theory. In *The Handbook of Stress and Health: A Guide to Research and Practice*; Wiley, **2017**, 349–364.
12. Cohen, S.; Kamarck, T.; Mermelstein, R. A Global Measure of Perceived Stress. *J. Health Soc. Behav*. **1983**, *24* (4), 385–396.
13. Kiker, D. S.; Motowidlo, S. J. Main and Interaction Effects of Task and Contextual Performance on Supervisory Reward Decisions. J. *Appl. Psychol*. **1999**, *84* (4), 602-609.
14. Borman, W. C. Expanding the Criterion Domain to Include Elements of Contextual Performance. In *Personnel Selection in Organizations*, Jossey-Bass, **1993**.
15. Halbesleben, J. R.; Wheeler, A. R. The Relative Roles of Engagement and Embeddedness in Predicting Job Performance and Intention to Leave. *Work Stress*. **2008**, *22* (3), 242–256.
16. Ursin, H.; Eriksen, H. R. Cognitive Activation Theory of Stress (CATS). *Neurosci. Biobehav. Rev*. **2010**, *34* (6), 877–881.
17. Maglio, P. P.; Campbell, C. S. Tradeoffs in Displaying Peripheral Information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*; ACM: New York, **2000**, 241–248.
18. Bailey, B. P.; Konstan, J. A.; Carlis, J. V. The Effects of Interruptions on Task Performance, Annoyance, and Anxiety in the User Interface. *Interact*; **2001**, 1, 593–601.
19. Hotchkiss, J. T.; Cook-Cottone, C. P. Validation of the Mindful Self-care Scale (MSCS) and Development of the Brief-MSCS among Hospice and Healthcare Professionals: a Confirmatory Factor Analysis Approach to Validation. *Palliat. Support Care*. **2019**, *17* (6), 628-636.
20. Wartig, S.L.; Forshaw, M. J.; South, J.; White, A. K. New, Normative English-sample Data for the Short Form Perceived Stress Scale (PSS-4). *J. Health Psychol*. **2013**, *18* (12), 1617-1628.
21. Van Dyne, L.; LePine, J. A. Helping and Voice Extra-role Behaviors: Evidence of Construct and Predictive Validity. *Acad. Manage J*. **1998**, *41* (1), 108-119.
22. O'Connor, D. B.; Shimizu, M. Sense of Personal Control, Stress and Coping Style: A Cross-Cultural Study. *Stress Health*. **2002**, *18* (4), 173–183.
23. Cox, T. Ferguson, E. Individual Differences, Stress and Coping. In Personality and Stress: Individual Differences in the Stress Process; Cooper, C. L., Payne, R., Eds.; John Wiley & Sons, **1991**, 7–30.

Authors

Megan Chung is a high school junior attending Seoul Foreign School in South Korea. She is interested in majors such as psychology, biology, and business, and is intrigued by how she can combine those majors for her future.

Gina Hwang is a high school sophomore attending UWC Southeast Asia, Singapore. She likes to understand the world using numbers and is interested in majors such as mathematics, economics, and business.

DOI: 10.36838/v7i8.14

IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Predicting Animal Population Trends Using Random Forest Models to Enhance Biodiversity Conservation

Aiden Chee

Taipei American School, 800 Zhongshan North Road, Section 6, Taipei, Taiwan, ROC 11152; aidenchee123@gmail.com
Mentor: Jimin Choi

ABSTRACT: Biodiversity is vital for ecological balance, as every species serves a specific function within its ecosystem. The rapid decline in certain animal populations highlights the urgent need for conservation efforts. This study employs a Random Forest model with an integrated data pipeline to predict animal population changes based on physical traits over time. Using the Living Planet Index (LPI) and cross-referenced Wikipedia data, the study examines features such as thermoregulation, habitat, diet, reproductive strategy, and flight capability. Missing data was addressed through forward filling, ensuring continuous and reliable datasets. Results show a minimal correlation between physical traits like habitat and thermoregulation and population trends, indicating that while physical traits offer insights, incorporating environmental or behavioral data is essential for accurate predictions. Future research can build on this framework by integrating advanced modeling techniques and broader datasets to improve biodiversity conservation strategies.

KEYWORDS: Animal Sciences, Ecology, Population Dynamics, Species Extinction, Ecological Modeling.

Introduction

This research aims to predict future animal population trends by analyzing physical traits such as habitat, thermoregulation, diet, and reproductive strategies. These traits influence species' adaptability to environmental changes. For example, climate shifts may impact endothermic animals differently than ectothermic ones, and species with specialized habitat needs may struggle as environments deteriorate. By identifying species at higher risk of population declines, this study seeks to guide conservation efforts toward proactive intervention.

Biodiversity is essential, as every species contributes to its ecosystem. Secondary and tertiary consumers regulate prey populations, pollinators like bees and hummingbirds support plant reproduction, and decomposers such as flies and isopods recycle nutrients. In 2023, 21 animal species were declared extinct in the United States, with estimates predicting up to 150 species will vanish globally each day, equivalent to one extinction every 10 minutes.[1] The United Nations warns that nearly 1 million species are at risk of extinction due to human activities.[2] Freshwater species populations have declined by 83% on average since 1970,[3] and the Amazon rainforest lost approximately 12,000 square kilometers of forest in 2022, an area comparable to Qatar.[3] Population declines disrupt ecosystems, as species' ecological roles go unfulfilled. For instance, gray wolves' extinction in Yellowstone National Park caused an unchecked rise in wild elk populations, leading to overgrazing and vegetation decline.[4] This example underscores how losing a single species can destabilize an entire ecosystem. Preserving current animal populations is, therefore, critical.

By examining the relationship between physical traits and population trends, this research identifies species more vulnerable to environmental changes based on their traits. It seeks to determine how physical traits correlate with population trends and whether these traits can predict species most at risk of population decline. This approach not only enhances understanding of how traits influence survival but also prioritizes conservation for species at higher risk, increasing the chances of preserving biodiversity before irreversible damage occurs.

Researchers argue whether traits alone can reliably predict extinction risk without considering environmental context.[5] This study contributes to that discussion by evaluating how well physical traits predict animal population trends when modeled with statistical population features. The use of a Random Forest regressor aligns with applications of machine learning in conservation, where similar models have been used to assess extinction risk, forecast species distributions, and identify vulnerability patterns across taxonomic groups.[6] By focusing on interpretable models and measurable traits, this study helps solidify the role of trait-based prediction in biodiversity risk.

While prior studies done by Qi have used Random Forest algorithms in bioinformatic settings to evaluate the importance of input features, they often do so in statistical contexts without direct usage in ecological population modelling.[7] Similarly, Moretti and Legg used plant and animal traits to assess responses to ecological disturbances, but did not use historical data.[8] This study extends both studies by using long-term population trends with ecological and physical attributes to better assess the predictive value of traits. Using feature importance values has been common when conducting Random Forest-based studies. However, our usage of the ecological setting could bring challenges in trait interpretation due to environmental variability. These studies provide a strong foundation for applying similar methods to predict species populations and assess vulnerabilities.

We have considered using other models, such as individual-based models (IBMs), which focus on behavior and

DOI: 10.36838/v7i8.15

physiology to predict species responses to environmental changes, making them useful for complex ecosystems. We also considered using dynamic range models (DRMs), which account for population movements and growth, and offer superior predictions in dynamic climates. While these approaches address specific aspects of species responses or general prediction techniques, they do not focus on analyzing physical traits to predict population trends across a broad range of species. This study builds on existing work by integrating physical traits and environmental factors to enhance prediction accuracy and inform targeted conservation strategies.

An efficient data pipeline is essential for this study, as it ensures smooth data flow from collection to model training and performance evaluation. Unlike typical conventional studies, this research deals with high-dimensional, scattered data, incorporating diverse physical and environmental factors across many species. Managing such complexity requires a robust pipeline to handle missing values, select features, and optimize models. This ensures data integrity and enhances prediction accuracy in trait-based population modeling. The pipeline supports data collection, preprocessing, model training, and evaluation.

One major challenge addressed was the prevalence of missing data in the population records. While the Living Planet Index (LPI) provides comprehensive data on over 32,000 populations and 5,200 species, its records for some species over 50 years are incomplete.[9] The pipeline resolved this issue by preprocessing the data, including imputing missing values to ensure continuity and completeness. This step was critical for preparing reliable datasets for modeling. This study utilized the LPI as a primary source due to its extensive biodiversity data. Preprocessing involved cleaning the data and handling missing values to establish a high-quality input for the model. The study sought to uncover correlations between the two by examining species populations alongside their physical traits. The Random Forest model was chosen for its ability to identify correlations and generate accurate predictions. Preprocessed data was fed into the model, which employs multiple decision trees. Each tree acts as an individual model, learning correlations between features to improve predictive accuracy. The model's performance was rigorously evaluated to identify areas for improvement and ensure reliable outcomes.

Methods

Data Collection and Sources:

We analyzed population trends using datasets from the Living Planet Index (LPI), developed by the Zoological Society of London (ZSL) and the World Wildlife Fund (WWF) (WWF/ZSL, 2022). This index tracks population changes across more than 32,000 populations and over 5,200 species, including mammals, birds, amphibians, reptiles, and fish. It gathers data from peer-reviewed studies, government reports, and wildlife surveys, offering a comprehensive view of how environmental changes affect species populations. Integrating the LPI dataset allowed us to examine historical population trends, identify correlations, and improve predictions for future patterns.

Table 1: Formatting of information in the Living Planet Index (LPI) dataset, including data on species populations, geographic locations, and time-series information. The table presents the structure of the dataset used for analysis.

Binomial	Biome	1950	…	1991	…	2020
Falco_punctatus	Tropical and subtropical…	NULL	…	22	…	NULL
Falco_punctatus	Atlantic north temperate	NULL	…	776000	…	1103000

Table 1 summarizes the LPI dataset, which provides taxonomic groupings and population data but lacks information on the physical traits of animals. We developed a data pipeline to fill this gap by sourcing additional details from external resources. Specifically, we used Wikipedia to extract and analyze the physical traits of various species. Recognizing the potential limitations of Wikipedia's credibility, we implemented a validation process. To ensure accuracy, this involved cross-referencing data with reliable scientific databases, such as the Global Biodiversity Information Facility. Using taxonomic classifications, we categorized animals based on shared physical traits, minimizing errors from relying on a single source.

We utilized the Wikipedia-API Python library to retrieve page content and identify relevant details through targeted keywords. Although Wikipedia served as the primary data source, our pipeline is adaptable for incorporating information from other credible databases in future research. This process enabled us to organize species into five sub-datasets, focusing on key traits: thermoregulation, habitat, dietary habits, flight capability, and reproductive strategies. These features were chosen for their influence on population trends. Thermoregulation affects metabolic rates and survival strategies, while habitat provides insight into environmental pressures on species. Dietary habits (e.g., carnivorous, omnivorous, herbivorous) reflect resource availability and feeding behavior. Reproductive strategies (e.g., live birth, egg-laying) influence growth rates and survival. Flight capability impacts species mobility and adaptation to environmental changes. These traits offer a comprehensive understanding of the ecological and biological factors shaping population dynamics.

Data Pipeline:

The effective management and processing of data is of great importance, as the analysis involves complex datasets with diverse features. The data we are working with spans many species and features, resulting in high-dimensional data that may complicate analysis. Additionally, combining data from various sources, such as the Living Planet Dataset (LPD_2022.csv) and the Wikipedia API, requires careful handling to ensure the integrity of the sources while being consistent and accurate. Furthermore, the presence of missing population data in the time-series form creates a challenge for creating continuity and reliability between data points.

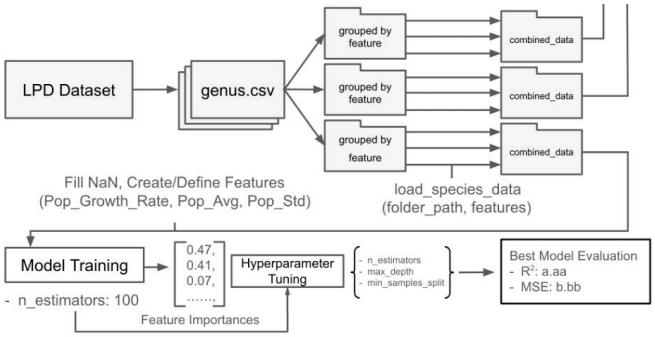


Figure 1: Overview of our data pipeline, illustrating the steps in data collection, preprocessing, and analysis. This pipeline was designed to ensure efficient integration of data into the predictive model, enabling reliable population trend analysis.

As shown in Figure 1, the initial phase of the data pipeline involves data collection, using the LPD dataset and the Wikipedia API. This stage establishes the foundation for subsequent analysis. Therefore, careful consideration must be given to selecting appropriate data sources and methods to ensure reliability.

To prepare the information for use in the Random Forest, the thousands of species needed to be grouped by features. This was addressed by first loading species data, which involved creating a function that would read CSV files from a specified directory. Each file contained data relating to a specific genus. Using the organized formatting of the files, important information such as species names, population counts over the years, and other relevant data was extracted.

The function "load_species_data" aggregated this information, where each dataset was transformed to a melted format, which would simplify analysis by merging data into columns for species, year, and population. Additionally, each entry was supplemented with details about its 'Blood_type' (thermoregulation), habitat, genus, and more, which were used in later stages of analysis.

This data preparation step was also supported by integrating information from the Wikipedia API. Augmenting the datasets with features such as thermoregulation and habitat enhanced the data further. The cross-referencing provided context for the numerical data and allowed the comprehensive analysis of the present ecological patterns. After loading and processing the data from the categorized species, these datasets were merged with species with similar groupings, and the resulting DataFrame was termed 'combined_data.' The merging process enabled the conduct of analyses spanning multiple species.

Data Cleaning and Imputation:

To ensure the quality of the data, missing values were addressed through strategies such as filling in missing data. The forward fill imputation was used to handle missing values in the population. This method propagates the most recent observed value forward to replace subsequent missing values, which is especially suitable for time-series data. By carrying forward the last known population value, the forward fill imputation is able to maintain continuity in the dataset without introducing completely unrealistic values, making it effective for the context.

We chose forward filling for this task to maintain continuity for species with many gaps in time-series data, however, the method may create bias if earlier values are not representative of later trends. Alternative techniques such as linear interpolation or KNN-based imputation were considered, but were not implemented due to data sparsity and computational cost. A comparative study of imputation methods could be a valuable direction for refinement in the future.

This preprocessing was essential as it prepared the data for subsequent statistical analyses and applications in machine learning. Furthermore, additional features, such as population growth rates, averages, and standard deviation, were added. These values were calculated for each species in the dataset, with $G = \frac{P_t - P_{t-1}}{P_{t-1}} \times 100$ represented population growth rate, where P_t and P_{t-1} represent the population in year t and the previous year t-1, respectively. The average population across all available years was calculated by $P_{avg} = \frac{1}{n}\Sigma_{i=1}^{n} P_i$, where n is the total number of years for which population data is available, and Pi is the population in year i. Finally, the standard deviation was calculated with the equation:

$$\sigma = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(P_i - P_{avg}\right)^2}$$

Feature Engineering and Final Dataset:

Additionally, the data pipeline used feature engineering to create new variables based on existing data. For instance, calculating derived metrics such as population growth rates, averages, and standard deviations from the raw data counts enriches the dataset and provides additional context that may improve model performance. This was added in the hopes that the model's ability to capture patterns and relationships in the data would be enhanced. This approach helps the Random Forest model use a more comprehensive set of features, which leads to more accurate predictions and insights about species populations and ecological relationships.

Ultimately, the data collection and preparation approach, which spanned from the initial loading of species data to the handling of missing values, created a foundation for further analysis using the Random Forest model. At the same time, feature engineering enabled us to make more accurate predictions and insights.

Methods

Random Forest Model:

The Random Forest model was selected for its ability to manage complex, high-dimensional data commonly encountered in ecological research. Unlike linear regression models, which assume linear trends, Random Forest captures non-linear associations with traits and trends, which is essential for ecological data. For example, relationships between environmental factors and species populations are rarely linear and can involve intricate interactions that linear models would miss. While linear methods may be sufficient in other contexts, the complexity of ecological data makes Random Forest better suited for predicting population trends.

DOI: 10.36838/v7i8.15

Random Forest achieves this by using an ensemble of decision trees, each trained on random subsets of the data. This ensemble approach minimizes overfitting, a common issue with single decision trees, and enhances the model's robustness to variations in the data. Moreover, Random Forest is resistant to outliers, as errors from individual trees tend to offset one another when aggregated. These attributes make it highly effective for analyzing real-world ecological data.

Another key advantage is its interpretability. The feature importance metric highlights the traits most influential in shaping population trends, enabling targeted conservation strategies, a central goal of this study. Although techniques like neural networks and support vector machines (SVMs) are viable alternatives, they require extensive tuning and often lack the interpretability that Random Forest provides. Although techniques like neural networks and support vector machines (SVMs) are viable alternatives, they require extensive tuning and often lack the interpretability that Random Forest provides. Neural networks are well-suited for identifying patterns in large, unstructured datasets, and SVMs excel in high-dimensional spaces, but their limitations in transparency make them less ideal for this project.

In contrast, Random Forest ideally balances predictive power and interpretability for understanding the drivers of population trends based on ecological factors. For an analysis where the relationships between variables and species populations are complex and non-linear, Random Forest is the best fit. While other techniques may offer benefits in certain contexts, the strengths of Random Forest align most closely with the goals of this study.

Model Inputs and Feature Selection:

This process enabled the extraction of meaningful insights from the data. In this study, the initial step involved selecting features and defining the target variable. The features included physical characteristics, habitat, population growth rate, standard deviation, and population averages, while the target variable was population counts. These inputs allowed the model to learn the relationships required for accurate predictions. An 80-20 train-test split ensured sufficient data for training and evaluation.

Data Preprocessing and Standardization:

Before fitting the model, features were standardized using the StandardScaler from scikit-learn, which scales them to have a mean of 0 and a unit variance. It transforms each feature x according to the using $z = \frac{x-\mu}{\sigma}$, where z is the standardized value, x is the original value of the feature, is the mean of the feature, and σ is the standard deviation of the feature. Using the StandardScaler reduced the impact of any discrepancies in scale among features.

Hyperparameter Tuning:

The first iteration of the Random Forest Regressor used 100 trees and was trained on the scaled training dataset. To enhance model performance even further, hyperparameter tuning was executed using GridSearchCV, an approach that evaluates a range of different combinations of hyperparameters.

GridSearchCV is especially beneficial when optimizing model performance, as it allows for thorough exploration of hyperparameters. By specifying a grid of hyperparameters, it is able to automate the tuning process, which ensures that all potential configurations are considered. It begins with defining the parameter grid and continues to employ cross-validation to evaluate the model's performance for each combination of parameters. This involves separating the dataset into K subsets, where the model is trained on K-1 folds. This process is repeated for every combination of parameters, allowing for an accurate assessment of each configuration's performance. Finally, it aggregates the results, calculating the mean performance metrics and identifying the best set of hyperparameters.

The tuned hyperparameters included the number of trees, maximum tree depth, and minimum samples required to split a node. For optimization, the parameter grid specified ranges of [100, 200, 300] for tree count, [10, 20, 30] for maximum depth, and [2, 5, 10] for minimum samples to split a node. This process identified a configuration that maximized predictive accuracy while maintaining resistance to overfitting.

Table 2: 3D table displaying the combinations of hyperparameters used in the analysis. The optimal combination found by tuning is bolded for clarity, ensuring reproducibility of future results.

Number of samples	Number of trees	Tree depth limit = 10	Tree depth limit = 20	Tree depth limit = 30
2	100	100, 10, 2	100, 20, 2	100, 30, 2
	200	200, 10, 2	200, 20, 2	200, 30, 2
	300	300, 10, 2	300, 20, 2	300, 30, 2
5	100	100, 10, 5	100, 20, 5	100, 30, 5
	200	200, 10, 5	200, 20, 5	200, 30, 5
	300	300, 10, 5	**300, 20, 5**	300, 30, 5
10	100	100, 10, 10	100, 20, 10	100, 30, 10
	200	200, 10, 10	200, 20, 10	200, 30, 10
	300	300, 10, 10	300, 20, 10	300, 30, 10

Model Evaluation and Metrics:

Following the identification of optimal parameters depicted in Table 2, the model with the highest R^2, a measure of how well the model accounts for differences in observed data, was evaluated on the test set from the train-test split to evaluate its predictive capabilities. Performance metrics such as R^2 and Mean Squared Error (MSE) were calculated. R^2 is measured with $R^2 = \frac{SS_{res}}{SS_{tot}} = \frac{\Sigma(y_i - y_{pred})^2}{\Sigma(y_i - y_{mean})^2}$, where SS_{res} is the sum of squares of the differences between the real and predicted values, and SS_{tot} is the total sum of squares, or the difference between the actual values and the mean of the values.

MSE is measured with $\Sigma(y_i - y_{pred})^2$, where n is the number of observations, y_i are the actual values, and y_{pred} are the predicted values. Lower MSE values illustrate better model performance.

In addition to numerical metrics, a visualization of prediction errors was conducted through the generation of a confusion matrix, shown in Figure 2. The confusion matrix summarizes predictions across multiple categories, with rows representing actual categories and columns representing the predicted categories. The diagonal cells show correct predictions for each category, while the cells outside represent misclassifications.

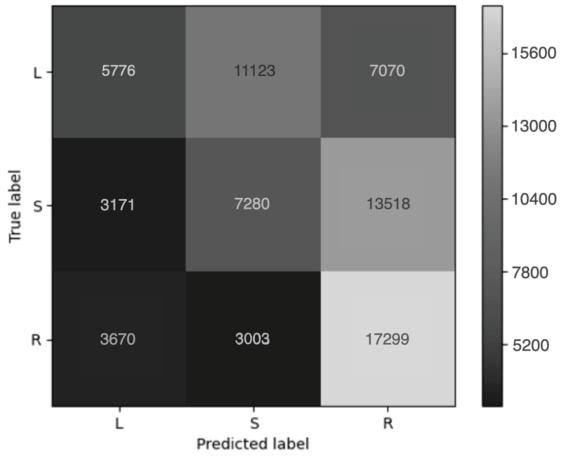


Figure 2: Confusion matrix showing the comparison of predicted and actual population values, categorized into three groups: Low (L), Medium (S), and High (R). This classification provides an alternative assessment of the model's performance in predicting population trends. Although correct predictions are present, the overall predictive power of the model is quite weak.

Results

The comparison of thermoregulation type, habitat, diet, reproductive strategy, and flight capability against statistical data such as species growth rate, average population across years, and standard deviation revealed minimal correlation with animal population trends.

The random forest regressor generates a Relative Importance Value (RIV). Calculated $RIV = \frac{\text{Importance}_{feature}}{\text{Importance}_{all}} \times 100$, the RIV value represents the importance of each feature to the predictions made by the model. This model calculates the RIV by measuring Gini impurity, which calculates feature importance and determines the final value. Gini impurity measures the likelihood of misclassifying a randomly chosen data point if labeled according to the class distribution. A feature's importance score is calculated based on its ability to decrease misclassification and improve decision-making in individual trees. In Random Forest, the RIV represents the average contribution of a feature to the reduction in Mean Squared Error (MSE) across all decision trees. Each split in the forest evaluates how much it reduces the MSE, and the features that result in greater reductions are assigned higher RIVs. The final RIV is the average reduction in error attributed to a feature across all trees, providing a robust measure of its importance in the model.

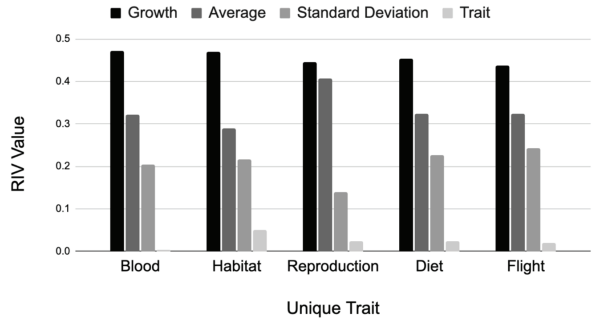


Figure 3: Relative Importance Values (RIV) for assessing the significance of five physical traits in predicting population trends. The RIV values are compared with statistical data to determine which traits most strongly correlate with population decline. The figure depicts low values for traits when compared to statistical data.

From the Random Forest regressor output, shown in Figure 3, the species' overall growth rate emerged as the most important feature in predicting future populations. It consistently held the highest RIV value compared to the other five features, indicating that species with stable or changing population trends are more likely to maintain these trajectories. Conservation efforts should prioritize species with declining populations, as they are at a higher risk of extinction. The standard deviation of populations over the years was then consistently the second most important feature. A higher standard deviation suggests greater fluctuations in population numbers, potentially signaling vulnerability to environmental changes or pressures.

The stronger performance of statistical features is likely since they encompass cumulative ecological effects over time. Growth rate and standard deviation reflect actual demographic responses to diverse pressures, such as habitat degradation, climate variability, or competition, without needing to explicitly model those factors. Physical traits, on the other hand, are more general, which means they encompass characteristics that may differently influence population trends under specific ecological conditions. As a result, while traits like thermoregulation and habitat are biologically meaningful, their predictive power is limited without environmental context, explaining why statistics outperformed trait-based variables.

In contrast, population growth rate, habitat, and thermoregulation features showed negligible RIV values, ranging from 0 to 0.05. This suggests that habitat changes may affect individual species rather than causing consistent changes across multiple species. The minimal RIV for thermoregulation indicates no significant correlation with future population trends.

The relatively low RIV value for habitat indicates that habitat changes do not consistently correlate with population changes across multiple species. Instead, these changes often alter species dynamics. This variability complicates the use of habitat as a reliable predictor for individual species populations, as seen in the application of Random Forest Regressor models. A relevant example is the extinction of gray wolves in Yellowstone National Park. The loss of this apex predator caused an unchecked increase in the elk population, which led to severe overgrazing. This vegetation loss degraded habitats, negatively impacting other species reliant on those ecosystems, illustrating the cascading effects of predator-prey dynamics on broader ecological systems.

The negligible RIV for thermoregulation indicates that population trends are likely shaped by combinations of factors rather than broad classifications alone. Simplistic classification risks overlooking significant variations in behavioral patterns, ecological adaptations, and physiological responses within these categories. For example, reptiles and fish may be classified together based on thermoregulation, but their population trends diverge due to significant differences in behavior and environmental sensitivities. Reptiles, such as lizards and snakes, often regulate body temperature by basking, influencing their activity and habitat use. Conversely, fish adjust their depth to manage temperature but are more vulnerable to water quality and temperature fluctuations, which directly affect

DOI: 10.36838/v7i8.15

breeding cycles and health. While reptiles are particularly susceptible to deforestation and its impact on food sources, fish face greater risks from aquatic environmental changes. These distinctions illustrate how population trends within thermoregulation-based groups can vary significantly. Relying solely on thermoregulation categories to assess populations may overlook critical factors such as competition for resources, predation pressures, and other environmental stressors. Species-specific traits and their interactions with complex ecological variables often provide more accurate insights into population dynamics than broad classifications.

Further analysis, however, reveals insights beyond just feature importance. First, the RIV values for certain features alone, such as habitat, do not exclude the possibility of interactions between variables. Although one feature alone may show minimal predictive influence, it could be more significant when combined with other features. For example, certain species in specific habitats may experience population changes due to environmental or ecological pressures that are not apparent when examining these features individually. Random Forest models can capture such interactions through decision tree splits. These splits divide data based on specific features or values, helping the model detect patterns. For example, a split could divide the data depending on whether the habitat type is "Marine." However, understanding these interactions requires more dedicated analysis techniques, including interaction effect plots or pairwise feature importance metrics, which assess the combined effect of two features. Using these tools could help clarify how features work together to influence predictions, which would create in-depth insights and likely inform conservation strategies that could address multiple factors simultaneously. This would provide a more comprehensive understanding of population patterns, especially for species that may be in complex environments.

While feature importance values provide interpretability in model behavior, we acknowledge their limitations in capturing causal relationships. An ablation study, where models are trained by removing features incrementally, was considered but not conducted due to the sparsity in our dataset and computational constraints. Future studies could expand on this study by the usage of ablation methods to evaluate feature combinations more accurately.

Regarding model accuracy, the stability of feature rankings across multiple Random Forest runs suggests consistent RIVs. This credibility further emphasizes features such as growth rate and average populations as important predictors. However, it is important to recognize any limitations in the application of RIVs.

Random Forests may favor features with more unique values, which could inflate their importance. To mitigate this, techniques like permutation importance can be used, which shuffle feature values to assess their true impact on prediction accuracy. Shuffling feature values could help in avoiding bias for abnormal values in the dataset, which would otherwise impact the importance values of features. From a conservation perspective, the findings suggest focusing on species with historically high population fluctuations, as the standard deviation indicates. These species may be more susceptible to minor environmental pressures. The high importance of population growth suggests that conservation efforts should target populations with declining trends, as these are likely to continue without intervention.

Discussion

The interconnectedness of species creates a complex web of interactions, making it challenging to discern overall population trends. Each species occupies a unique ecological role and responds differently to environmental pressures. For example, while large herbivores may thrive in the absence of predators, competing species or those dependent on vegetation face adverse effects, such as habitat degradation or reduced food availability. Understanding whether a species' population is increasing or declining requires a broader ecological context, as factors like competition, mutualism, and environmental changes significantly influence responses to habitat shifts.

The inclusion of historical data aimed to establish a baseline for understanding population trends, but it should not dominate the analysis when predicting future changes. The primary focus is on analyzing how physical traits, such as thermoregulation and habitat, influence species' adaptability and survival in dynamic environments. While historical population sizes provide a reliable foundation for estimating future sizes, they are less informative for identifying changes in population trends—the core objective of this study. The results suggest that excluding historical data in future iterations of the model may enhance its alignment with the study's goals. By focusing on the impacts of physical traits and emphasizing small deviations as early indicators for conservation efforts, the model can more effectively predict population trends and support targeted interventions.

Limitations of Physical Traits:

Although physical traits certainly influence a species' ability to survive in different conditions, they do not alone account for the shaping of population dynamics. For instance, while a group of species may have a consistent diet or reproductive strategy, such as being herbivorous or laying eggs, other aspects of the animals may be extremely varied due to other physical traits of the animal or different habitats. In this sense, the reasoning behind the low predictive scores for thermoregulation is also reflected when measuring the scores of flight capability, diet, and ways of birthing offspring. Each of these traits plays a role within a broader web of factors that influence population trends. For instance, while flight capability is crucial for certain species adapting to environmental changes, it does not address how populations are impacted by predator-prey dynamics or resource availability. Similarly, other traits often interact with environmental and ecological factors in ways that minimize or alter their individual influence on population trends. Thermoregulation affects sensitivity to climate variability, where ectothermic species are more susceptible to extreme temperatures. These examples illustrate how the predictive power of individual traits can be limited without considering their combined effects. Exploring such combinations could reveal

hidden vulnerabilities that are not apparent when examining traits in isolation, where methods like pairwise interaction analysis within Random Forest models could provide deeper insight into how physical traits shape population resilience or decline. This highlights the need to analyze a more comprehensive set of variables, emphasizing interactions among traits and environmental factors. Considering these combinations can provide a more accurate understanding of population dynamics and their broader ecological implications.

Implications:

The findings of this research highlight the importance of refining analyses to better understand how physical traits interact with environmental factors and population dynamics. Given the complexity of population trends, future studies should explore how traits like habitat interact with historical population data to improve the accuracy of predictions. Monitoring changes in population trends can help identify species at risk due to climate change, insights that may not be apparent from historical data alone. While historical population data is useful for predicting trends in stable conditions, physical traits become critical for understanding changes under increasing environmental pressures.

Future Research and Model Expansion:

This study also introduced methodological advancements. The data processing pipeline effectively managed extensive missing values, and trait data were sourced using the Wikipedia API and taxonomic groupings. These developments create an infrastructure for further exploration of similar topics, potentially supporting more effective and stronger analyses in future studies. Additionally, the flexibility of the Random Forest model allows for the integration of new features, traits, or population data to expand the model's range. Future research could incorporate environmental indicators, such as climate change or pollution data, to capture ecological interactions influencing population trends within specific groupings. The use of more detailed subcategories based on alternative or mixed physical and behavioral traits could further enhance the precision of population analyses.

Enhancing the pipeline with automated processes for feature selection and hyperparameter tuning by using grid search or other evolutionary algorithms could also optimize model performance as well as prevent the model from overfitting. Another potential improvement involves exploring ensemble methods, combining Random Forests with models like Dynamic Range Models (DRMs) or Individual-Based Models (IBMs). These hybrid approaches could better capture overall population trends while accounting for specific behavioral differences at the individual level. Continued development of this framework holds promise for creating more effective predictive tools. By incorporating advanced methodologies and diverse modeling approaches, future research can support more accurate analyses, guiding conservation efforts and species protection initiatives with greater precision.

Real Time Monitoring Applications:

The predictive framework developed in this study could be used to support real-time monitoring systems for conservation decision making. By using continuously updated population data from monitoring programs to feed directly into this model, population counts processed through the Random Forest algorithm would pre-emptively detect population decline based on species-specific traits. This would enable conservation managers to identify species with abnormal declines and prioritize interventions before substantial loss occurs. By automating this process, the model could support a real-time alert system for regions or species where consistent population tracking is available.

■ Conclusion

This study assessed the role of physical traits in predicting changes in animal populations, focusing on thermoregulation, habitat, diet, reproductive strategy, and flight capability. The results show that historical population data, including growth rate, standard deviation, and average population size, were crucial predictors of population trends. However, physical traits provided valuable insight into population changes, where small deviations in population trajectories can lead to significant ecological shifts. The Random Forest model demonstrated that the historical population data effectively predicted trends based on past patterns. While physical traits showed lower Relative Importance Values (RIVs) individually, their interactions with other environmental and biological factors may reveal more complex relationships. For instance, habitat may exert a stronger influence on population trends when analyzed alongside additional ecological variables, emphasizing the importance of assessing these interactions.

Although historical data is valuable for estimating population levels, the physical traits studied hold promise for identifying deviations from ongoing trends. Such deviations, even minor ones, can serve as early indicators of larger ecological or environmental changes, offering insights into species' long-term survival. Predicting changes in population trends requires a broader framework that integrates historical data with various physical and ecological factors for a more comprehensive understanding.

■ Acknowledgments

I would like to thank my mentor, as well as Dr. Matthew Caesar, a Computer Science professor at the University of Illinois Urbana-Champaign, for providing insights during the writing process and planning stages. Their guidance allowed me to shape my ideas into a more cohesive, well-structured paper.

■ References

1. Kaufman, M. These animals went extinct in 2023. *Mashable*, 2023, December 27. https://mashable.com/article/extinct-species-animals-2023 (accessed Dec 27, 2023).
2. Bongaarts, J. IPBES, 2019. Summary for Policymakers of the Global Assessment Report on Biodiversity and Ecosystem Services of

DOI: 10.36838/v7i8.15

the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. *Population and Development Review*. 2019, *45* (3), 680–681. https://doi.org/10.1111/padr.12283.
3. An 83% decline in freshwater animals underscores the need to keep rivers connected and flowing. *World Wildlife Fund*. https://www.worldwildlife.org/stories/an-83-decline-of-freshwater-animals-underscores-the-need-to-keep-rivers-connected-and-flowing (accessed Dec 27, 2023).
4. White, P. J.; Garrott, R. A. Northern Yellowstone elk after wolf restoration. *Wildlife Society Bulletin* 2005, *33* (3), 942–955. https://doi.org/10.2193/0091-7648(2005)33.
5. Pacifici, Michela, et al. "Assessing Species Vulnerability to Climate Change." Nature Climate Change, vol. 5, 2015, pp. 215–224. https://doi.org/10.1038/nclimate2448.
6. Cardillo, Marcel, et al. "Multiple Causes of High Extinction Risk in Large Mammal Species." Science, vol. 309, no. 5738, 2005, pp. 1239–1241.
7. Qi, Y. Random Forest for bioinformatics. In *Springer eBooks*; Springer: New York, 2012; pp 307–323. https://doi.org/10.1007/978-1-4419-9326-7_11.
8. Moretti, M.; Legg, C. Combining plant and animal traits to assess community functional responses to disturbance. *Ecography* 2009, *32* (2), 299–309. https://doi.org/10.1111/j.1600-0587.2008.05524.x.
9. Almond, R. E. A.; Grooten, M.; Juffe Bignoli, D.; Petersen, T. (Eds.) *Living Planet Report 2022 – Building a nature-positive society*; WWF: 2022. https://wwf.panda.org (accessed Dec 27, 2023).
10. Johnston, A. S. A.; Boyd, R. J.; Watson, J. W.; Paul, A.; Evans, L. C.; Gardner, E. L.; Boult, V. L. Predicting population responses to environmental change from individual-level mechanisms: towards a standardized mechanistic approach. *Proceedings of the Royal Society B: Biological Sciences* 2019, *286* (1913), 20191916. https://doi.org/10.1098/rspb.2019.1916.
11. Pagel, J.; Schurr, F. M. Forecasting species ranges by statistical estimation of ecological niches and spatial population dynamics. *Global Ecology and Biogeography* 2011, *21* (2), 293–304. https://doi.org/10.1111/j.1466-8238.2011.00663.x.

Authors

Aiden Chee:

Aiden Chee is a tenth-grade student at Taipei American School. He is interested in environmental sciences as well as engineering and is passionate about preventing further damage within ecosystems due to a lack of awareness.

Jimin Choi:

University of Michigan, Ann Arbor, USA

IJHSR International Journal of High School Research

is a publication of

terra science and education

N.Y. based 501.c.3 non-profit organization
dedicated for improving K-16 education