IJHSR

International Journal of High School Research

January 2026 I Volume 8 I Issue 2

ijhsr.terrajournals.org

ISSN (Print) 2642-1046

ISSN (Online) 2642-1054

GENIUS OLYMPIAD

Let's build a better future together

International Project Fair focused on Sustainability and Environment for Grades 8-12

- Since 2011
- Hosted around 1400 participants in 2025 from 35+ states 70+ countries
- **Disciplines:** STEM, Coding, Robotics, AI, Speech, Entrepreneurship, Arts, Short Film, Music
- Applications start on December 1
- Application Deadline is March 1
- Finalists are announced by March 25
- Event is usually scheduled 2nd week of June
- Monday – Friday, includes a trip to Niagara Falls
- Hosted by large universities at Upstate New York
- Application Fee is $60/ project
- Participation Fee is $600/ person, w/ room and board
- Open buffet breakfast, lunch, and dinner
- Trip to Niagara Falls and boat tour is included
- **Instagram and Facebook @Geniusolympiad**
- For more information: **GENIUSOlympiad.org**
- Email: **info@geniusolympiad.org**

GENIUS Olympiad is organized by Terra Science and Education, a N.Y. based 501.c.3 non-profit organization dedicated for project-based learning

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

January 2026| Volume 8| Issue 2

Editorial Board

International Journal of High School Research

IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

An Integrated Lock-in Amplifier and Near Infrared Method for Finding Fruits Fully Behind Leaves in a Homemade Testbed

Aiden Xu

Winter Springs High School, 130 Tuskawilla Rd, Winter Springs, FL 32708, USA; goaidenxu@gmail.com

ABSTRACT: Detecting fully occluded objects is of interest for various practical problems, such as harvesting and yield prediction in farming, which are physically demanding and heavily labor-dependent. Many approaches have been explored by researchers aiming to solve this problem. However, they are ineffective due to inherent challenges: the strength of signals reflected from hidden objects is weak, and those signals are always buried in high-magnitude noise. In this study, a method combining near-infrared (NIR) and lock-in-amplifier (LIA) techniques is proposed to tackle these challenges. Two questions are answered. Can a fully covered fruit be detected purely based on reflected NIR signals? Can LIA extract reflected signals from high-magnitude noise? This study addresses these questions from theoretical and experimental points of view, including NIR photon particle propagation, LIA in the image format, low-cost experiment apparatus, etc. In total, 268 videos were collected over 134 valid experiments with tomatoes and cucumbers as objects. Both alternate hypotheses were validated and answered.

KEYWORDS: Embedded Systems, Sensors, Occluded Fruit Detection, Near Infrared, Lock-in-Amplifier.

Introduction

Finding what's behind or hidden in leaves is a key step in many applications. For example, many farming activities are labor-intensive and physically demanding, such as yield prediction, leaf thinning, harvesting, and pesticide applications.[1-4] Among them, harvesting is mostly done manually,[4] especially for fruit crops like tomatoes, cucumbers, and strawberries. However, labor is in short supply in the US,[5] which means more robots are needed. For a robot to effectively conduct those tasks currently done by humans, it needs to know if there is something (e.g., fruit, flower, or peduncle) behind dense leaves.

In the past decade, many researchers have investigated different methods to solve the aforementioned problems. Most of them utilized vision-based, artificial intelligence (AI) methods.[1, 6-12] A method to detect tomatoes using visible light cameras and machine learning was investigated as well.[12] Another study used a leaf blower to mechanically expose hidden apples so a LIDAR could be used more effectively to detect them.[1] However, to date, none of them have been highly successful. The main issues are: (i) the reflected signal from hidden fruits is weak, and (2) the reflected signal is buried in high magnitude noise. The author also noticed that, very recently, researchers [13, 14] used millimeter wave radar techniques in finding fruits behind leaves[14] with relatively higher cost, lower reflectivity on soft material surfaces, and the need for a specialized imaging system.

In this study, a method combining near-infrared (NIR) and lock-in-amplifier (LIA) [15] in the image format is proposed to address these issues. There are two sets of hypotheses. In Hypothesis Set 1, "effective" means the method is effective in detecting the presence of an object fully hidden behind leaves. "Scenario 1" represents a scenario with a fully occluded object, while "Scenario 2" represents a scenario without such an object. In Hypothesis Set 2, "effective" means the proposed method is better than the simple image subtraction method (the control group) in detecting an object.

Null Hypothesis 1 (N1): If in more than 30% of the experiments, the reflected NIR signal in "Scenario 1" is NOT significantly different from that of "Scenario 2", then the proposed method is NOT "effective".

Alternative Hypothesis (AH1): If in more than 70% of the experiments, the reflected NIR signal in "Scenario 1" is significantly higher than that of "Scenario 2", then the proposed method is "effective".

Null Hypothesis 2 (N2): If in more than 30% of the experiments, the percentage difference of the LIA technique is NOT higher than the simple subtraction method, then the proposed method is not "effective".

Alternative Hypothesis 2 (AH2): If in more than 70% of the experiments, the percentage difference of the LIA technique is higher than that of the simple subtraction method, then the proposed method is "effective".

The research conducted to validate those hypotheses consists of three main parts. The first part is to select the diodes with the best wavelength considering cost, product availability, and optical properties on leaves and fruits. The second part is to create an innovative, in-house testbed: the "emitter" box (producing NIR signals modulated with the Pulse Width Modulation - PWM), the "orchard" box (housing leaves and fruit), and the "phone holder" (a stable base for a cell phone to detect NIR signals and record experiments). The third part, LIA in the image format, is the most innovative one. Software for signal generation and data analysis was also developed.

The contributions of this study are as follows. As far as the author knows, there are two technical contributions. (i) The test apparatus can conduct experiments to validate the research hypotheses despite costing much less than any optical equipment in research laboratories. (ii) It is the first time a combined

DOI: 10.36838/v8i2.1

technology of NIR and LIA has been tried in detecting fully occluded fruits. On a broader scale, this research has the potential to reduce labor dependence and enable more efficient robotic operations in harvesting, yield prediction, etc. If combined with different electromagnetic waves, this research can benefit an even wider range of applications, e.g., robot motion in off-road environments and medical imaging,[16] etc.

The paper is organized as follows. Firstly, I will discuss the theoretical background, test apparatus, data analysis tools, and experiments. Then, the experimental data and findings will be shown. Discussions, limitations, and conclusions are given in the end.

■ Methods

Theoretical Background:

1. NIR photon particle propagation and detection

Figure 1 shows the sketch of how the NIR photon particles propagate in the custom-designed experiment apparatus (discussed later). *D*, I_o, A_{out}, and A_{in} represent the detector efficiency, initial NIR intensity, signal attenuation outside of leaves, and signal attenuation inside of leaves. The leaf transmissivity, object reflectability, and leaf reflectability are denoted by L_T, O_R, and L_R, respectively.[2] $I_{F_{in}}$ and $I_{F_{out}}$ are the reflected NIR signal intensity detected by the camera.

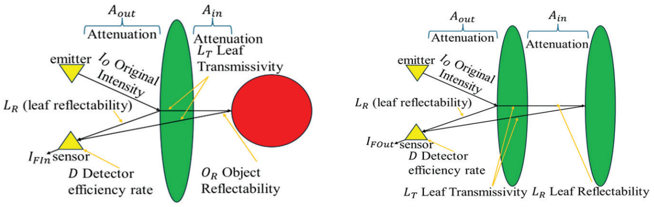

A_{out}
Attenuation
A_{in}
Attenuation
L_T Leaf Transmissivity
emitter
I_O Original Intensity
L_R (leaf reflectability)
I_{FIn} sensor
D Detector efficiency rate
O_R Object Reflectability
A_{out}
Attenuation
A_{in}
Attenuation
emitter
I_O Original Intensity
L_R (leaf reflectability)
I_{FOut} sensor
D Detector efficiency rate
L_T Leaf Transmissivity
L_R Leaf Reflectability

Figure 1: NIR signal propagation. (Left) with a fruit fully behind leaves and (right) no fruit behind. In this experiment setup, the PWM-modulated NIR signals are emitted, and a camera or detector receives the reflected NIR signals. There are two paths for the reflected signals: (i) directly reflected by the leaves and (ii) transmitted through leaves and reflected by the fruit (left) or leaves (right). Based on this experiment sketch, an equation can be derived to determine the difference in reflected signal intensities between the cases with and without a hidden fruit.

The NIR signal strength when a fruit is behind leaves, $I_{F_{in}}$, is derived as

$$I_{FIn} = D\left[\left(I_O A_{out}^2 L_T{}^2 A_{in}^2 O_R\right) + \left(I_O A_{out}^2 L_R\right)\right] \tag{1}$$

which considers the NIR signal reflected directly from the leaves and the NIR transmitted through leaves, bounced back from the hidden fruits, and then transmitted through the leaves again.

Similarly, the NIR signal strength when there is no fruit behind leaves, $I_{F_{out}}$, is derived as

$$I_{FOut} = D\left[\left(I_O A_{out}^2 L_T{}^2 A_{in}^2 L_R\right) + \left(I_O A_{out}^2 L_R\right)\right] \tag{2}$$

Therefore, the difference between $I_{F_{out}}$ and $I_{F_{in}}$, represented by ΔI, is derived as

$$\Delta I = I_{FIn} - I_{FOut} = D I_O L_T^2 A_{out}^2 A_{in}^2 (O_R - L_R) \tag{3}$$

One way to increase ΔI is to increase the initial intensity I_o; therefore, 20 diodes are used based on the test apparatus volume. Secondly, the wavelength with a high L_T, low L_R, and a high O_R should be chosen. Based on the optical experiment, the Gikfun® 940nm diodes were adopted (also low cost).

2. Lock-in amplifier in the image format

The LIA technique has been widely used to extract useful but weak signals buried from large magnitude of noise that with frequencies different from the reference signal.[15] Figure 2 shows how the LIA method is customized in the image format for the custom-designed experiment. In the scenarios of fruits being fully hidden behind leaves, as shown in Eq. 3, the reflected NIR signal differences between the scenarios with and without hidden fruits are very small.

The Arduino instructs the NIR diodes to emit a signal A_I modulated with a PWM square wave in its Fourier series $\sum_{i=1}^{n} m_i \sin[(2i-1)\omega t + \phi]$.[17] Here, A_I can be $I_{F_{in}}$ (Eq. 1) or $I_{F_{out}}$ (Eq. 2), ω is the foundational frequency, *t* is the time, and ϕ is the phase angle.[17] $m_i = 4/[\pi(2i-1)]$,*i*=*1*, …, *n*, is the coefficient in the Fourier series expansion with n harmonics.[17] The detected signal S_I is

$$S_I = \{A_I \textstyle\sum_{i=1}^{n} m_i \sin[(2i-1)\omega t + \phi] + N_I\} \tag{4}$$

where N_I is noise (e.g., random, specific frequency). As shown in Figure 2, the signal S_I goes through an average filter to remove random noise, and is then multiplied by a reference signal *R* (PWM) to output signal S_O. After that the signal SO goes through a Butterworth low-pass filter (LPF),[18] and the remaining DC component is S_{OL}. As shown in the derivation in Appendix A, A_I equals S_{OL}. The equations used in the custom designed experiment are shown in Appendix A. The process of using LIA is illustrated in Figure 2.

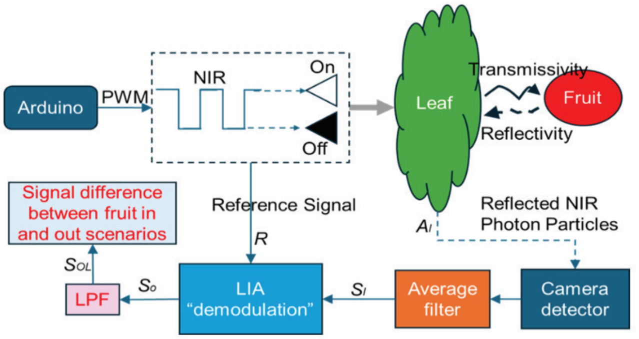


Figure 2: Signal flow chart in the experiments and data analysis. The NIR diodes will emit NIR signals which are modulated by PWM. The camera will detect reflected NIR signals. An average filter is first used to remove random noise, and then the resultant signal modulated with the reference signal through the LIA demodulation. After an LPF, the reflected NIR signal is calculated. As shown in later experiments, this method is effective in validating the alternative hypotheses.

Test Apparatus:

The test apparatus went through five design iterations, and only the final version is shown here.

1. NIR emitter box design:

As shown in Figure 3-left side, the "emitter" box produces a NIR signal modulated with PWM, has an access point to the Arduino and an external button to control signal starting, holds all necessary circuitry, and reflects minimal light to reduce noise. The "emitter" box is based in a 25.38x17.77x17.77 cm^3 wooden box, and 20 holes were drilled on one side for the diodes. The circuits are controlled by an Arduino Mega®. A button is present for controlling the signal's start. Each of the 4 breadboards connects with five diodes. The diodes are arranged in two circles (Figure 3-right side). The inner circle has 8 diodes, and the outer has 12 diodes. This pattern was determined by considering the limitations coming from size and volume constraints of the emitter box, diodes, and wires. To minimize reflected light, all exterior surfaces except the back were painted black. On each breadboard, one side hosts two diodes and the other hosts three. For the side with two, since each diode requires 1.2V and the Arduino outputs 5V, 2.6V is taken by the resistor. Since the diode's working current is 30mA, an 87Ω resistor is needed for that part of the circuit. Following a similar calculation, the resistor used in the 3-diode circuit is 47Ω. Since the legs of a diode were too short to reach the breadboard, soldering jumper cables is required.

2. Orchard box design and phone holder:

The "orchard" box must house leaves and fruit and keep them in their spots during an experiment, as well as minimize light reflection. Thus, the "orchard" box, shown in Figure 3 (left side), has three horizontal lines of string across the front. The topmost is where leaves are attached; the other two prevent the leaves from curling inward. Behind them is a raised platform, where the object is placed. The orchard box inside is covered in black foam to minimize light reflecting off it.

The "phone holder" needs to provide a low-cost, stable base for a cell phone to detect reflected NIR signals and record experiments. As such, it is built of plastic building bricks. It is hollow in the middle, for holding and steadying the cell phone to keep it in the same place while recording in different experiments. The cell phone's brightness is set to the minimum to avoid emitting excess light.

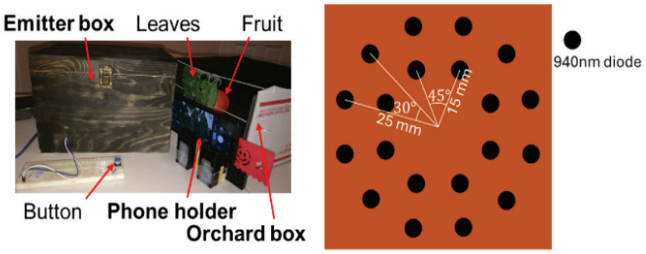


Figure 3: Test apparatus (left) and the layout of diodes (right). On the left, the test apparatus includes an "emitter" box for NIR signal generation, an "orchard" box holding leaves and fruit, and a phone holder to support the camera. On the right, there are 20 NIR diodes arranged in two concentric circles to increase the signal intensity. As a result, the apparatus is efficient and low-cost, and can be easily made from materials found around the household.

Data Analysis and Software:

About 1,600 lines of code (six codes) were programmed in Arduino® and MATLAB®.

1. NIR signal modulated with PWM:

(Code 1) The Arduino® code is to instruct the diodes to emit NIR signals modulated with PWM (6 seconds or 10 seconds, with 10 periods for each experiment). The signal is turned on by pressing a button to sync signal generation with video recording.

2. Data analysis tools:

(Code 2) Before the data analysis tools are applied, three signals (fruit in and out NIR signals and the PWM reference signal) should be synchronized. The data retrieval code extracts the RGB values of pixels and takes each frame's average RGB values, acting as an average filter.

(Code 3) The first data analysis method is the simple subtraction method, serving as the "control" group. This method simply subtracts the image without a fruit from the image with a fruit. Code 4 and Code 5 are for the LIA and the LIA with a Butterworth LPF.[18] Code 6 is to implement a dual LIA method with LPF, and interested readers can find how a dual LPF works.[15]

Experiments:

1. Leaf and fruit optical properties experiment:

The optical property experiment was conducted at a University of Central Florida laboratory using an Evolution 220® spectrophotometer following the procedure in Figure 4. According to the experiment results and following Eq. (3), the 940nm wavelength diodes were selected.

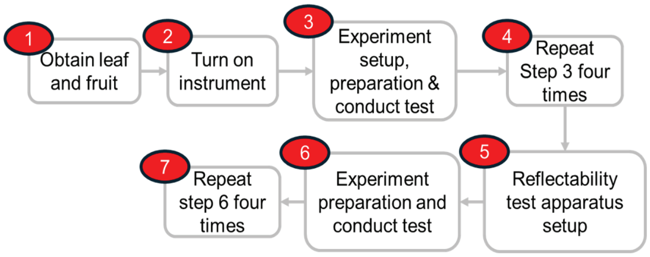

1
Obtain leaf and fruit
2
Turn on instrument
3
Experiment setup, preparation & conduct test
4
Repeat Step 3 four times
5
Reflectability test apparatus setup
6
Experiment preparation and conduct test
7
Repeat step 6 four times

Figure 4: Procedure of leaf and fruit optical property experiment. The procedure follows the guideline of the instrument, and experiments were conducted to check the reflectivity and transmissivity of fruit and its corresponding leaf. It was found that the 940nm wavelength diode would be ideal, because (i) for leaves it has relatively low absorption and reflectability, and high transmissivity, and (ii) for fruits it has low absorption and transmissivity and high reflectability, in addition to being low-cost.

2. Experiments of detecting fully covered objects:

As shown in Figure 5, first, obtain enough leaves to cover the front of the "orchard" box and a fruit. The leaves are taped to the topmost string, and the fruit is placed on the platform. Next, the "emitter", "orchard", and "phone holder" are arranged properly, with 2.54 cm or 0 cm of distance between the "emitter" and "orchard" boxes, with the "phone holder" wedged between the two. The computer is then connected to the Arduino, and the PWM signal period is set to either 6 or 10 seconds. Both the record button and the signal start button

DOI: 10.36838/v8i2.1

are pressed at the same time to start. Once 10 periods are over, stop the phone recording. Now, remove the fruit and repeat the process. Each experiment consists of two scenarios: one with an object fully covered by leaves and the other without such an object; and this is counted as one independent replicate.

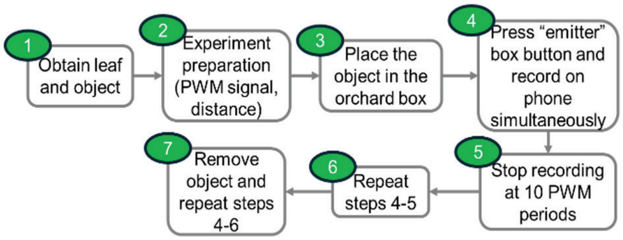

1
Obtain leaf and object
2
Experiment preparation (PWM signal, distance)
3
Place the object in the orchard box
4
Press "emitter" box button and record on phone simultaneously
5
Stop recording at 10 PWM periods
6
Repeat steps 4-5
7
Remove object and repeat steps 4-6

Figure 5: Experiment procedure of detecting fully occluded fruits. As shown in the results below, the experiment procedure is effective at validating the alternate hypotheses.

Result and Discussion

Experiment Data:

A total of 178 experiments were conducted over 20 weeks, and 356 videos were collected. However, not all of them were used, as some were invalidated due to an experiment setup error causing high amounts of ambient noise, while incorrect types of leaves were used in other invalidated experiments. In the end, 134 experiments and their 268 videos were used in the data analysis.

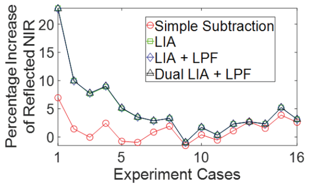
Experiment Results:

Table 1 shows the number of experiments that fulfill the requirements of AH1 (a), AH2 (b), and both (c), respectively, in the format of (a, b, c). For example, (36, 37, 34) represents the number of experiments using tomato fruit that validated AH1, AH2, and both, respectively. In 107 out of 134 experiments, both alternate hypotheses are validated (Table 1).

Table 1: Successful experiments in validating the alternate hypotheses. Experiments were conducted for four settings: PWM periods (6s or 10s) and the distance between "emitter" and "orchard" (0cm or 1" (2.54cm)). A total of 134 experiments are shown here. The number of experiments that can validate AH1, AH2, and both are listed in the form of (a, b, c), respectively. Both AH1 and AH2 are supported because the percentages of the successful detections are above 70%.

Experiments (AH1, AH2, Both)					
Scenarios	1" 6s	1" 10 s	0" 6s	0" 10s	Total
Tomato	(11, 11, 10)	(8,8,8)	(10, 10, 9)	(7,8,7)	(36, 37,34)
Tomato peduncle	(6, 6, 6)	(6, 6, 6)	(4, 6,4)	(6, 6, 6)	(22, 24, 22)
Cucumber	(9,10,7)	(22, 15, 14)	(12, 8, 8)	(14, 14, 14)	(57, 47, 43)
Cucumber peduncle	(2, 2, 2)	(2, 2, 2)	(2, 2, 2)	(4, 2, 2)	(10, 8, 8)

The following figures show the detailed experiment results of different fruits and different experiment configurations. In Figures 6 and 7, it is obvious that the LIA methods extracted significantly higher signals as compared to the simple subtraction method when a fruit is there. However, the difference between when peduncles are there or not is not obvious (Figure 8).



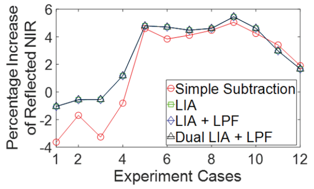


Figure 6: % increase of reflected NIR (2.54 cm distance, 6s period) with a tomato fruit. As compared with the control group (using the simple subtraction method), the LIA methods have larger percentage increases when there is a fully hidden fruit.

Figure 7: % increase of reflected NIR (2.54 cm distance, 6s period) with a cucumber. Similar findings are found as in Figure 6.

Figures 9 and 10 show the overall detection rates of different fruits and peduncles. The detection rate for tomatoes is above 87% (Figure 10) as compared to above 48% in the control group (Figure 9), signifying that the proposed method is more effective. In addition, since the peduncle detection rates are below 20%, the proposed method can differentiate between fruits and peduncles.

Statistical tools are used to analyze results. In Figures 11 and 12, the mean values in both the control and experiment groups are positive and mostly above 1% when detecting hidden fruits, meaning AH1 is supported. In addition, the mean value bars are located higher when using the LIA method as opposed to the simple subtraction method, supporting AH2. Those observations are not obvious when peduncles are used, meaning the proposed method can tell the difference between fruits and peduncles. The trend in standard deviation values in those figures is similar for both control and experiment groups. However, that is because in this custom designed experiment scenario, the majority of noise is random noise, which is filtered out by an average filter used in both SSM and LIA methods. Thus, their standard deviation trends are similar. However, since the LIA method can remove noise with frequencies different from the reference signal, its results are slightly better, and thus AH2 is supported.

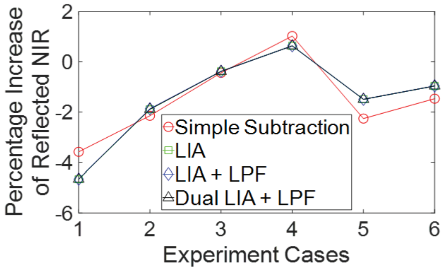


Figure 8: % increase of reflected NIR (2.54 cm distance, 6s period) with tomato peduncles. Null hypotheses are supported because the percentage difference between the control group (the simple subtraction method) and the proposed LIA method is not significant. However, this is as expected, since it means that the method can differentiate between fruits and peduncles.

DOI: 10.36838/v8i2.1

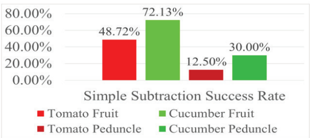


Figure 9: The overall success rate of the simple subtraction method. The detection rate in hidden fruit cases is significantly higher than those of peduncle cases.

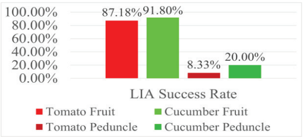


Figure 10: The overall success rate of the LIA method. The detection rate when using the proposed LIA method is much higher than those of the simple subtraction method, which validates alternate hypothesis 2.

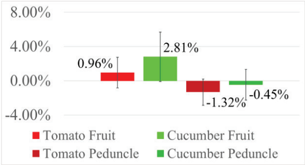


Figure 11: Mean +/- standard deviation percentage increase of reflected NIR (simple subtraction). The mean values of signal percentage increases are 0.96% and 2.81% for tomato and cucumber cases, respectively, meaning AH1 is supported.

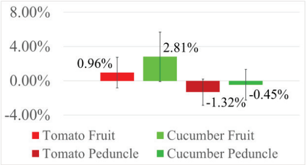
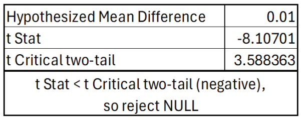


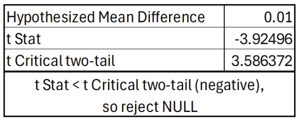
Figure 12: Mean +/- standard deviation percentage increase of reflected NIR (LIA). As compared with the control group, the proposal LIA method can achieve a much higher signal percentage increase.

The following figures show the t-test between the simple subtraction and LIA methods. In both Figure 13 and Figure 14, the t-stat is less than the negative t-critical two-tail, meaning that LIA is better than the simple subtraction method, rejecting N2 and supporting AH2.



Hypothesized Mean Difference	0.01
t Stat	-8.10701
t Critical two-tail	3.588363
t Stat < t Critical two-tail (negative), so reject NULL	

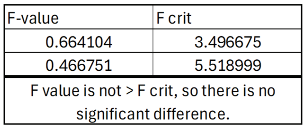
Figure 13: T-test for the simple subtraction and LIA methods (tomato). Here, the t-stat value (-8.10701) is less than -3.588363, meaning N2 is rejected and AH2 is supported.



Hypothesized Mean Difference	0.01
t Stat	-3.92496
t Critical two-tail	3.586372
t Stat < t Critical two-tail (negative), so reject NULL	

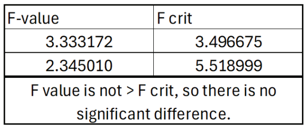
Figure 14: T-test for the simple subtraction and LIA methods (cucumber). Here, the tstat number (-3.92496) is less than the negative t-critical two tail value (- 3.586372). Therefore, AH2 is supported and N2 is rejected.

The following two figures (Figure 15 and Figure 16) show the ANOVA tests between different experiment configurations: namely, 2.54 cm-6s, 2.54 cm-10s, 0 cm-6s, and 0 cm-10s, for the simple subtraction and LIA methods. In both figures, the F value is not larger than the F critical value, so there is no statistical difference. This means that the small distances and periods do not have a major effect on the performance of the proposed method.



F-value	F crit
0.664104	3.496675
0.466751	5.518999
F value is not > F crit, so there is no significant difference.	

Figure 15: ANOVA test for the differing distance and period for simple subtraction. The results mean that minor distances and differences in the period of the PWM signal do not have a major effect on the accuracy in the control group.



F-value	F crit
3.333172	3.496675
2.345010	5.518999
F value is not > F crit, so there is no significant difference.	

Figure 16: ANOVA test for the differing distance and period for the LIA method. The results mean that minor distances and differences in the period of the PWM signal do not have a major effect on the accuracy in the proposed LIA method.

Discussions, Limitations, and Future Work:

According to the results, the reflected NIR signal in "Scenario 1" is higher than that of "Scenario 2", as the total difference in percentage between them is 2.67%, so AH1 is validated. In most experiments, the difference in the reflected NIR signals is more prominent when using the LIA method as opposed to the simple subtraction method, as the difference percentage for the LIA method is 3.46% compared to the simple subtraction method of 1.88%, so AH2 is validated.

However, there are some limitations. (i) The "orchard" box cannot completely imitate actual conditions. Future work includes adding more layers of leaves. In addition, fruits and peduncles could be shown at the same time. (ii) In the current experiment setup, ambient light is not fully blocked, which may cause minor errors, which can be addressed by adding a band-pass filter. (iii) Due to the sub-optimal quality of the current camera, further investigation into a better NIR detector will be conducted. (iv) Three statistical analysis methods, those being mean/standard deviation, t-test, and ANOVA test, are used in this study. More statistical methods will be used for comprehensive analyses in the future work. (v) In this study, only the SSM method is considered in the control group; and in the future, other information processing method could be investigated and compared with the proposed LIA method.

Conclusion

This research studies a combined NIR and LIA method to detect fruits fully hidden behind leaves. A very low-cost test apparatus was designed and built, using which 134 valid experiments were conducted, yielding 268 videos. Both AH1 and AH2 are supported by experiment data. The t-test shows that the proposed LIA method is effective in detecting fully occluded fruits than the SSM method. This research can significantly enhance farming operations' efficiency, such as in harvesting and yield prediction.

Acknowledgments

The author thanks his main research advisor, Dr. Shawn Putnam, for suggesting and explaining the method of LIA and NIR, and how it can potentially be used to detect hidden fruits, Dr. Yajie Dong for providing access to spectrophotometer equipment, and Dr. Yunjun Xu for allowing the author to use the soldering station, providing the financial support, and proofreading the manuscript.

Appendix A

The LIA equations relating to a sinusoid reference signal can be easily found in literatures.[15] The procedure in obtaining the LIA equation with a PWM reference signal is briefly explained here. As shown in Figure 2 (the custom designed experiment testbed), the signal A_I (with noise) goes through an average filter to remove random noise, which becomes S_I. Then it is multiplied by a reference signal R (PWM) as

$$S_O = S_I R = \{A_I \sum_{i=1}^{n} \frac{4}{\pi(2i-1)} \sin[(2i-1)\omega t + \phi] + N_I\}\{\sum_{i=1}^{n} \frac{4}{\pi(2i-1)} \sin[(2i-1)\omega t]\} \quad (A1)$$
$$= (\frac{4}{\pi})^2 A_I \sum_{n=1,3,5,\ldots}^{\infty} \sum_{m=1,3,5,\ldots}^{\infty} \frac{1}{mn} \sin(n\omega t) \sin(m\omega t) + \frac{4}{\pi} N(t) \sum_{m=1,3,5}^{\infty} \frac{1}{m} \sin m\omega t$$

DOI: 10.36838/v8i2.1

In Eq. (A1), the second term will be removed by a low pass filter to obtain its DC component S_{OL}. This DC component is the same as the reflected signal without noise, meaning $A_I=S_{OL}$. This result is well known; however, the detailed derivation seems not readily available in open literature. Interested readers may reach out to the author for the detailed derivation.

Note 1: The constant coefficient in $A_I=S_{OL}$ does not affect the arguments in the Section of "Results and Discussion," as the results are based solely on the ratio.

References

1. Gené-Mola, J., Ferrer-Ferrer, M., Gregorio, E., Blok, P. M., Hemming, J., Morros, J., Rosell-Polo, J. R., Vilaplana, V., Ruiz-Hidalgo, J., Looking behind occlusions: A study on a modal segmentation for robust on-tree apple fruit size estimation. *Computers and Electronics in Agriculture* **2023**, *209*, 107854. DOI: doi.org/10.1016/j.compag.2023.107854
2. Lu, R., Beers, R. V., Saeys, W., Li, C., Cen, H., Measurement of optical properties of fruits and vegetables: A review. *Postharvest Biology and Technology*, **2020**, *159*, 111003. DOI: doi.org/10.1016/j.postharvbio.2019.111003
3. Mishra R., Karimi D., Ehsani R., Lee W. S., Identification of citrus greening (HLB) using a VIS-NIR spectroscopy technique. *The ASABE Annual International Meeting*, **2012**, *55*, 711-720. DOI: doi.org/10.13031/2013.41369
4. Kusumastuti, R. D., Van Donk, D. P., Teunter, R., Crop-related harvesting and processing planning: a review. *International Journal of Production Economics*, **2016**, *174*, 76-92. DOI: doi.org/10.1016/j.ijpe.2016.01.010
5. *Farm Labor*; Economic Research Service, USDA, revised August 2025. https://www.ers.usda.gov/topics/farm-economy/farm-labor
6. Gongal, A., Amatya, S., Karkee, M., Zhang, Q., Lewis, K., Sensors and systems for fruit detection and localization: A review. *Computers and Electronics in Agriculture*, **2015**, *116*, 8-19 DOI: doi.org/10.1016/j.compag.2015.05.021
7. Choi, D., Lee, W. K., Schueller, J. K., Ehsani, R., Roka, F., Diamond, J., A performance comparison of RGB, NIR, and depth images in immature citrus detection using deep learning algorithms for yield prediction. *The ASABE Annual International Meeting*, **2017**, 1700076. DOI: doi:10.13031/aim.201700076
8. Mahmud, M. S., Zahid, A., He, L., Choi, D., Krawczyk, G., Zhu, H., LiDAR-sensed tree canopy correction in uneven terrain conditions using a sensor fusion approach for precision sprayers. *Computers and Electronics in Agriculture*, **2021**, 191, 106565. DOI: doi.org/10.1016/j.compag.2021.106565
9. Mirbod, O., Choi, D., Heinemann, P. H., Marini, R. P., He, L., On-tree apple fruit size estimation using stereo vision with deep learning-based occlusion handling. *Biosystems Engineering*, **2023**, 226, 27-42. DOI: doi.org/10.1016/j.biosystemseng.2022.12.008
10. Mirbod O., Choi, D., Thomas, R., He, L., Overcurrent-driven LEDs for consistent image colour and brightness in agricultural machine vision applications. *Computers and Electronics in Agriculture*, **2021**, 187, 106266. DOI: doi.org/10.1016/j.compag.2021.106266
11. Wang, J., He, D., Song, J., Dou, H., Du, W., Non-destructive measurement of chlorophyll in tomato leaves using spectral transmittance. *IJABE*, **2015**, 8, 73-78. DOI: http://doi.org/10.3965/j.ijabe.20150805.1931
12. Yamamoto, K., Guo, W., Yoshioka, Y., Ninomiya, S., On plant detection of intact tomato fruits using image analysis and machine learning methods. *Sensors*, **2014**, 14, 12191-12206. DOI: https://doi.org/10.3390/s140712191
13. Shi, Y., Ma, Y., Geng, L., Apple detection via near-field MIMO-SAR imaging: a multi-scale and context aware approach, *Sensors*, **2025**, 25, 1536. DOI: https://doi.org/10.3390/s25051536
14. Shiraz, Z., Khan, U. M., Shahzad, M., FruitSight: a millimeter wave radar based approach to detect occluded fruits, *2024 IEEE 21st International Conference on Mobile Ad-Hoc and Smart Systems*, **2024**. DOI: https://doi.org/10.1109/MASS62177.2024.00044
15. Low Level Optical Detection Using Lock-in Amplifier Techniques. *PerkinElmerTM Instruments*, 1–8. chem.ucla.edu.
16. Cuccia, D. J., Bevilacqua, F., Durkin, A. J., Tromberg, B. J., Modulated imaging: quantitative analysis and tomography of turbid media in the spatial-frequency domain. *Optics Letters*, **2005**, 30, 1354-1356. DOI: doi.org/10.1364/OL.30.001354
17. Mathworld. *Fourier Series — Square Wave*. https://mathworld.wolfram.com/FourierSeriesSquareWave.html
18. MATLAB, www.MathWorks.com, last accessed on 9/16/2024.

Authors

Aiden Xu is a high school student at the Winter Springs High School, Winter Springs, FL. He is interested in conducting different types of research during his spare time and has been participating in the Science and Engineering Fair since 6th grade.

IJHSR

ijhighschoolresearch.org

■ RESEARCH ARTICLE

Analysis and Discussion of Solutions to the Schrödinger Equation Under a Sawtooth Potential

Zhixuan, Tao

Westlake Boys High School, 30 Forrest Hill Road, Forrest Hill, Auckland, 0620, New Zealand; jason.tao.personal@gmail.com

ABSTRACT: Sawtooth potentials—piecewise-linear potentials defined by alternating rising and falling slopes—have garnered interest for their roles in classical Brownian ratchets, quantum transport, and flat bands in ultracold-atom lattices. However, a general solution of the time-independent Schrödinger equation for an arbitrary asymmetric sawtooth potential and a study of its band structure remain absent from the literature. We derive the exact analytical eigenstates in terms of Airy functions, but due to numerical difficulty, we solved for the eigenstates using a piecewise constant "staircase approximation." We find localized low-energy states confined within individual wells and plane waves for high energies. We apply the obtained solutions to investigate the relationship between bandwidth and parameters of the sawtooth potentials, which could offer design principles for realizing flatbands in condensed-matter and ultracold-atom research. We find that the bandwidth of the lowest energy band is entirely independent of the sawtooth's asymmetry, while it remains inversely correlated with potential height and cell size. The methodologies presented here provide a toolkit for further exploration of sawtooth potentials.

KEYWORDS: Physics and Astronomy, Theoretical, Computational and Quantum Physics, Solid State Physics, Sawtooth Potentials, Flat bands.

■ Introduction

Sawtooth potentials, characterized by a pattern of regions of increasing potential followed by regions of decreasing potential, have long been of interest across many fields. Sawtooth potentials are studied extensively in Brownian ratchet theory, where asymmetric sawtooth potentials are found to drive unidirectional transport when energy input is used to switch the potential between one of 2 states (a "flashing ratchet" that rectifies Brownian transport).[1] Such properties of the classical sawtooth potential are what drive motor proteins and many other essential intracellular transport processes.[2] Quantum mechanical treatment of sawtooth potentials also demonstrates transport phenomena. By exposing Bose-Einstein condensates to sawtooth-like optical lattices whose amplitudes are periodically modulated with time, a directed atomic current is observed despite the lack of dissipative processes.[3]

Sawtooth potentials are of interest in condensed matter physics for their ability to create flat bands.[4] A flat band is an energy band where energy is largely independent of the crystal momentum; this allows weak interactions to dominate. Flat bands are theorized to host a wealth of exotic behaviors, including high-T superconductivity, Wigner crystallization, and complex ferromagnetic behaviors.[5] Ultracold atoms in tunable optical lattices that have sawtooth characteristics have been utilized to engineer a nearly flat energy band in the sawtooth geometry.[4]

Much of the literature surrounding a quantum mechanical treatment of sawtooth potentials focuses on their ability to drive transport. A common area of study is to examine the transport properties of the sawtooth potential through the Schrödinger equation.[3-7] The exact solution to an asymmetric V-shaped potential has been used to investigate how the asymmetry of a sawtooth potential could affect tunneling probabilities and thus the transport properties of the sawtooth potential,[6] but the solutions to the Schrödinger equation for the sawtooth potential were never obtained. We will, in this paper, instead focus on obtaining solutions to the time-independent Schrödinger equation under a sawtooth potential, and examining its band structure, neither of which has been done for a generally applicable case in the literature. We hope the methods introduced in this paper will help others research into the theory behind the sawtooth potential, and that the findings may offer insight into design principles for flat bands.

■ Methods

Defining variables:

We first define the sawtooth potential by the following constants (Figure 1). We consider a single sawtooth to be a section of increasing potential, followed by a section of decreasing potential. Let $V_{\mathbf{top}}$ be the height of the potential at the tip of the sawtooth. Let the slope of the increasing side of the sawtooth be $\boldsymbol{k_1}$ and the decreasing side by $\boldsymbol{k_2}$. By definition, $\boldsymbol{k_1 > 0}$ and $\boldsymbol{k_2 < 0}$. The total width T of a sawtooth will then be $\frac{V_{top}}{k_1} - \frac{V_{top}}{k_2}$. The asymmetry $\boldsymbol{\xi}$ of a sawtooth potential will be given by the fraction of the sawtooth across which the potential is rising, given by the following $\xi = \frac{k_2}{k_2 - k_1}$, with $\mathbf{0 < \xi < 1}$.

DOI: 10.36838/v8i2.7

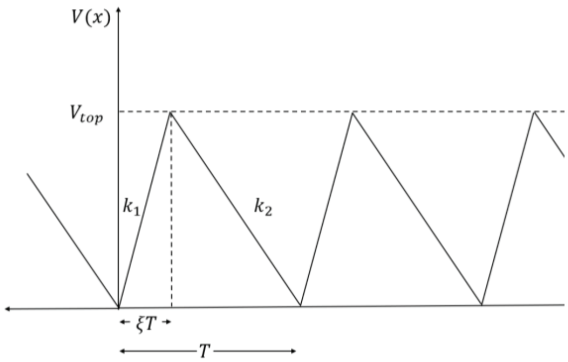


Figure 1: The parameters for our sawtooth potential. Across a singular sawtooth of size *T*, the potential increases linearly to V_{top} over a distance ξ *T*, then decreases linearly to 0 over a distance (1- ξ)*T*.

The eigenstates for a singular sawtooth:

Let's examine the rising potential side with slope k_1. This gives the potential function V(x) = k_1x. Plugging into the time-independent Schrödinger equation, we arrive at

$$\frac{\hbar^2}{2m}\frac{d^2}{dx^2}\psi(x) + xk_1\psi(x) = E\psi(x) \tag{1}$$

$$\frac{\hbar^2}{2m}\frac{d^2}{dx^2}\psi(x) + (xk_1 - E)\psi(x) = 0 \tag{2}$$

This differential equation is close to the form ***y" +xy* = 0**, whose solution is known to be a linear combination of Airy functions.[8] Our aim now will be to express the above differential equation in the form of ***y"+xy* = 0**. We first define a new variable $z_n = \left(\frac{2mk_n}{\hbar^2}\right)^{1/3}\left(x - \frac{E}{k_n}\right)$. For simplicity's sake, we will define a constant $a_n = \left(\frac{2mk_n}{\hbar^2}\right)^{1/3}$. Note that our potential will still be defined in terms of x. Rewriting the TISE in this variable with n=1, we arrive at the following.

$$z_1 = \mathrm{a}_1\left(x - \frac{E}{k_1}\right) \tag{3}$$

$$\frac{\hbar^2}{2m}\frac{d^2}{dx^2}\psi(z_1) + \frac{k_1 z_1}{a_1}\psi(z_1) = 0 \tag{4}$$

Applying the chain rule and the fact that z_n is linear in ***x***, we arrive at

$$\frac{d^2}{dx^2}\psi(z_1) = \left(\frac{dz_1}{dx}\right)^2\frac{d^2}{dz_1^2}\psi(z_1) = a_1^{\,2}\frac{d^2}{dz_1^2}\psi(z_1) \tag{5}$$

and as $\mathbf{a^3/k_1 = 2m/\hbar^2}$,

$$\frac{d^2}{dz_1^2}\psi(z_1) + z_1\psi(z_1) = 0 \tag{6}$$

The time-independent Schrodinger equation has been reduced to the form ***y"+xy* = 0**, as shown above, the solution to which is shown below. $C_{A,1}$ and $C_{B,1}$ are arbitrary constants, where the number subscript denotes the value of n, and the letter denotes whether its Ai(x) or Bi(x).

$$\psi(z_1) = C_{A,1}Ai(z_1) + C_{B,1}Bi(z_1) \tag{7}$$

Given that **Bi(*x*)** diverges as ***x***→∞ (Figure 2), $\mathbf{C_{B,1} = 0}$, as otherwise the wavefunction will not be normalizable: **ψ(*x*)** must vanish as ***x***→±∞. By plugging in k_2, we obtain the wavefunction for the 2 sides of the sawtooth. Furthermore, the wavefunction and its derivative must be continuous such that the second derivative of the wavefunction present in the TISE is defined. Hence, we enforce the following boundary conditions at ***x* = 0**.

$$C_{A,1}\mathrm{Ai}(-a_1E/k_1) = C_{A,2}\mathrm{Ai}(-a_2E/k_2) \tag{8}$$

$$\frac{C_{A,1}}{C_{A,2}} = \frac{\mathrm{Ai}(-a_2E/k_2)}{\mathrm{Ai}(-a_1E/k_1)} \tag{9}$$

(continuity of the wavefunction)

$$\frac{C_{A,1}}{C_{A,2}} = \frac{a_2\mathrm{Ai}'(-a_2E/k_2)}{a_1\mathrm{Ai}'(-a_1E/k_1)} \tag{10}$$

(continuity of the derivative)

This set of boundary conditions allows us to solve for the values of En in the spectrum of the wavefunction. However, given the non-elementary expression of the Airy functions, solving for the spectrum will need to be done numerically.

$$\psi(x)_n = \begin{cases} C_{A,2}\mathrm{Ai}(a_2(x - \frac{E_n}{k_2})) & x \leq 0 \\ C_{A,1}\mathrm{Ai}(a_1(x - \frac{E_n}{k_1})) & x \geq 0 \end{cases} \tag{11}$$

Normalization of a single sawtooth and solving for the exact values of ***C*** in this way can be done numerically. However, as the goal is to extrapolate this solution to an infinite periodic sawtooth potential, which is not normalizable over all **x**, we do not discuss these results further.

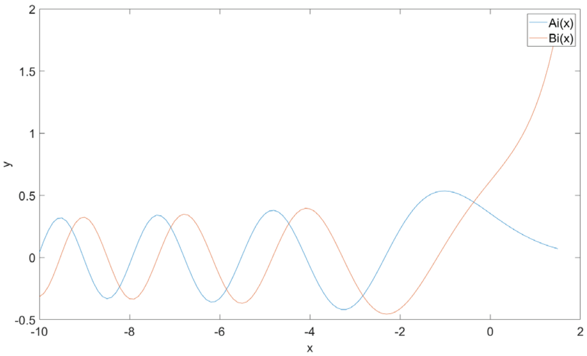


Figure 2: Ai(*x*) and Bi(*x*). Note Bi(*x*)'s divergence with increasing *x*.

The infinite periodic sawtooth potential:

When extrapolating our results from the previous section to an infinite periodic sawtooth, we may include **Bi(*x*)** terms as the function is never allowed to diverge to infinity, as each length of the slopes in a sawtooth is finite. In the unit cell at the origin, on the increasing slopes, we have the following wave function.

$$\psi_1(z_1) = C_{A,1}Ai(z_1) + C_{B,1}Bi(z_1) \tag{12}$$

DOI: 10.36838/v8i2.7

And on the decreasing slopes, the wavefunction is

$$\psi_2(z_2) = C_{A,2}Ai(z_2) + C_{B,2}Bi(z_2) \tag{13}$$

Where $\boldsymbol{C}_{\mathrm{A,1}}$, $\boldsymbol{C}_{\mathrm{B,1}}$, $\boldsymbol{C}_{\mathrm{A,2}}$, and $\boldsymbol{C}_{\mathrm{B,2}}$ are constants to be determined through boundary conditions and subsequent normalization. A reminder that $\boldsymbol{z}_{\mathbf{n}}=\left(\frac{2mk_n}{\hbar^2}\right)^{1/3}\left(x-\frac{E}{k_n}\right)$ and $a_n = \left(\frac{2mk_n}{\hbar^2}\right)^{1/3}$. The boundary conditions at x=0 are as follows

$$C_{A,1}\mathrm{Ai}(-a_1E/k_1) + C_{B,1}\mathrm{Bi}(-a_1E/k_1) \\ = C_{A,2}\mathrm{Ai}(-a_2E/k_2) + C_{B,2}\mathrm{Bi}(-a_2E/k_2) \tag{14}$$

$$a_1C_{A,1}\mathrm{Ai}'(-a_1E/k_1) + a_1C_{B,1}\mathrm{Bi}'(-a_1E/k_1) \\ = a_2C_{A,2}\mathrm{Ai}'(-a_2E/k_2) + a_2C_{B,2}\mathrm{Bi}'(-a_2E/k_2) \tag{15}$$

When employing boundary conditions for the points where 2 unit cells meet, we must apply Bloch's theorem, which states that solutions to the Schrödinger equation in a periodic potential can be expressed as plane waves modulated by periodic functions.[9] Essentially, for periodic potentials that span the entire x-axis, due to the translational symmetry of $|\boldsymbol{\psi}(\boldsymbol{x})|^2$, unit cells of the overall wavefunction may only differ by a phase $\mathbf{e}^{ikT}$ where T is the length of one unit cell.

$$\psi(x+T) = e^{ikT}\psi(x) \tag{16}$$

This can be expressed as a boundary condition on our wavefunction and is given below.

$$C_{A,1}\mathrm{Ai}(a_1(\frac{V_{top}}{k_1} - \frac{E}{k_1}) + C_{B,1}\mathrm{Bi}(\frac{V_{top}}{k_1} - \frac{E}{k_1}) \\ = e^{ikT}(C_{A,2}\mathrm{Ai}(\frac{V_{top}}{k_2} - \frac{E}{k_2}) + C_{B,2}\mathrm{Bi}(\frac{V_{top}}{k_2} - \frac{E}{k_2})) \tag{17}$$

$$C_{A,1}a_1\mathrm{Ai}'(a_1(\frac{V_{top}}{k_1} - \frac{E}{k_1}) + C_{B,1}a_1\mathrm{Bi}'(a_1(\frac{V_{top}}{k_1} - \frac{E}{k_1})) \\ e^{ikT}(C_{A,2}a_2\mathrm{Ai}(a_1(\frac{V_{top}}{k_2} - \frac{E}{k_2})) + C_{B,2}a_2\mathrm{Bi}(a_2(\frac{V_{top}}{k_2} - \frac{E}{k_2})) \tag{18}$$

We apply this to our wavefunction at points where $\mathbf{V}(\boldsymbol{x})=V_{\mathbf{top}}$. We note that this set of boundary conditions, after rearranging for 0, is a homogeneous system of linear equations where the variables are $\boldsymbol{C}_{\mathrm{A,1}}$, $\boldsymbol{C}_{\mathrm{B,1}}$, $\boldsymbol{C}_{\mathrm{A,2}}$, and $\boldsymbol{C}_{\mathrm{B,2}}$,. For such systems, there is either only a trivial solution ($\boldsymbol{C}_{\mathrm{A,1}}=\boldsymbol{C}_{\mathrm{B,1}}= \boldsymbol{C}_{\mathrm{A,2}}= \boldsymbol{C}_{\mathrm{B,2}}=\mathbf{0}$) or an infinite number of solutions, of which we are only interested in the latter. By rearranging Eq. 14-18 for 0, We express this system of equations in matrix form, where $\mathbf{M}$ represents the coefficient matrix for the aforementioned variables.

$$M = \begin{bmatrix} \mathrm{Ai}(\alpha) & \mathrm{Bi}(\alpha) & -\mathrm{Ai}(\beta) & -\mathrm{Bi}(\beta) \\ a_1\,\mathrm{Ai}'(\alpha) & a_1\,\mathrm{Bi}'(\alpha) & -a_2\,\mathrm{Ai}'(\beta) & -a_2\,\mathrm{Bi}'(\beta) \\ \mathrm{Ai}(\gamma) & \mathrm{Bi}(\gamma) & -e^{ikT}\,\mathrm{Ai}(\delta) & -e^{ikT}\,\mathrm{Bi}(\delta) \\ a_1\,\mathrm{Ai}'(\gamma) & a_1\,\mathrm{Bi}'(\gamma) & -e^{ikT}a_2\,\mathrm{Ai}'(\delta) & -e^{ikT}a_2\,\mathrm{Bi}'(\delta) \end{bmatrix} \begin{bmatrix} C_{A,1} \\ C_{B,1} \\ C_{A,2} \\ C_{B,2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\alpha = -\frac{a_1E}{k_1} \\ \beta = -\frac{a_2E}{k_2} \\ \gamma = a_1\left(\frac{V_{top}}{k_1} - \frac{E}{k_1}\right) = \frac{a_1(V_{top}-E)}{k_1} \\ \delta = a_2\left(\frac{V_{top}}{k_2} - \frac{E}{k_2}\right) = \frac{a_2(V_{top}-E)}{k_2} \tag{19}$$

The condition for an infinite number of solutions reduces to $\mathbf{det(M) = 0}$; we solve for the null space of $\mathbf{M}$.[10] The free variable we are left with can then be determined by normalization.

An approximation:

An exact solution for $\boldsymbol{C}_{\mathrm{A,1}}$, $\boldsymbol{C}_{\mathrm{B,1}}$, $\boldsymbol{C}_{\mathrm{A,2}}$, $\boldsymbol{C}_{\mathrm{B,2}}$ and $\boldsymbol{E}$ is difficult due to the many transcendental elements present inside $\mathbf{M}$. However, we can approximately solve for this potential by using an approximation. We employ a piecewise-constant approximation for our sawtooth potential. We discretize our potential into $\boldsymbol{N}$ equal-length "slices" of a constant potential (Figure 3). The value of the constant potential is the midpoint of the potential between the endpoints of the slice. The solution to this "staircase potential" asymptotically approaches the solution of the exact sawtooth potential as $\boldsymbol{N} \rightarrow \infty$. The oscillatory solutions to the Schrodinger equation for a constant potential are numerically more stable when one imposes boundary conditions: Airy

$$\frac{\hbar^2}{2m}\frac{d^2}{dx^2}\psi_n(x) + (V_n - E)\psi_n(x) = 0 \tag{20}$$

$$\frac{d^2}{dx^2}\psi_n(x) + \frac{2m(V_n-E)}{\hbar^2}\psi_n(x) = 0 \tag{21}$$

$$\kappa_n = \sqrt{\frac{2m(V_n-E)}{\hbar^2}} \quad \psi_n(x) = e^{\pm\kappa_n x} \tag{22}$$

$$\psi_n(x) = P_1\,cos\big(\kappa_n(x-n\varDelta x)\big) \\ + P_2\,sin\big(\kappa_n(x-n\varDelta x)\big) \tag{23}$$

$$\psi_n'(x) = P_1\kappa_n\,cos\big(\kappa_n(x-n\varDelta x)\big) \\ - P_2\,\kappa_n sin\big(\kappa_n(x-n\varDelta x)\big) \tag{24}$$

functions are oscillatory for $\boldsymbol{x} < \mathbf{0}$ and then rapidly decaying/growing for $\boldsymbol{x} > \mathbf{0}$ (Figure 2).

Here we represent the exponential solutions in a trigonometric basis (though for energies greater than the potential, they are hyperbolic as $\boldsymbol{\kappa}$ becomes imaginary). To determine the constants $\mathbf{P_1}$ and $\mathbf{P_2}$, we employ a transfer matrix method.[11] The transfer matrix $\mathbf{Q}$ that propagates across one of the slices is as follows.

DOI: 10.36838/v8i2.7

$$\begin{pmatrix}\psi_n(x+\Delta x)\\ \psi_n'(x+\Delta x)\end{pmatrix} = Q_n(x+\Delta x, x)\begin{pmatrix}\psi_n(x)\\ \psi_n'(x)\end{pmatrix} \tag{25}$$

$$Q_n(x+\Delta x, x) = \begin{pmatrix}\cos\big(\kappa_n(x-x_0)\big) & \dfrac{1}{\kappa_n}\sin\big(\kappa_n(x-x_0)\big)\\ -\kappa_n\sin\big(\kappa_n(x-x_0)\big) & \cos\big(\kappa_n(x-x_0)\big)\end{pmatrix} \tag{26}$$

We then multiply all transfer matrices for ***N*** slices to get the overall transfer matrix for one unit cell, **Q_{total}**. In this paper, ***N* = 500** was used. Enforcing the Bloch boundary conditions returns an eigenvalue equation for the total transfer matrix that then defines the allowed energies E and the dispersion relationship.[11] This can further be simplified to an equation in terms of **Tr(Q_{total})**.

$$\mathrm{Tr}\ (Q_{total}) = 2\cos(kT) \tag{27}$$

This equation was solved numerically using a root-finder algorithm in MATLAB. Reconstructing the wave function afterwards is straightforward. Note that one needs to choose the initial values for **ψ(*x*)** and **ψ'(*x*)** at some arbitrary x. However, because of the arbitrariness of the overall phase for the wavefunction and the subsequent unit cell normalization we perform, the exact initial choice does not matter so long as the initial transfer matrix is not singular.

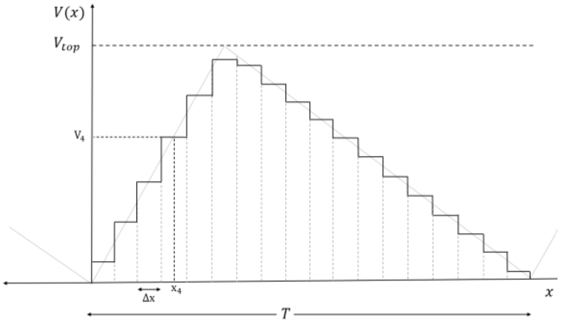


Figure 3: The staircase potential. *N*=18 was used here for effect: *N*=500 is used for actual results. The potential of the 4th slice V_4 is taken as the potential of the exact sawtooth potential at the midpoint of the slice x_4.

Result and Discussion

We used a piecewise constant approximation to obtain the band structure for a sawtooth potential of arbitrary asymmetry **ξ**, height ***V_{top}***, rising slope ***k_1***, and falling slope ***k_2*** (Figure 4). From this, we reconstructed the solutions to the time-independent Schrodinger equation under a sawtooth potential. Throughout all visualizations from here, we have taken mass to be the mass of an electron **(*m* = 9.11x10^{-31} kg)**. The band structures were visualized up to the first Brillouin zone for ***0* ≤ *k* ≤ *π/T*** sufficient to capture all the unique states due to the periodicity of the Bloch phase.

We will now examine the wavefunction and the band structure for **ξ = 0.3**, ***V_{top}* = 0.3eV**, and ***T* = 1nm**. Low energy solutions **(*E* < *V_{top}*)** are localized within each well of the sawtooth (Figure 5). The general profile of the wave function did not change significantly with energy for lower energies. In the low-energy regime, the probability density peaks at the lowest point of the sawtooth potential and dips throughout the sawtooth itself, demonstrating significant localization. When ***E* > *V_{top}***, the solutions (Figure 6 and Figure 7) resemble plane waves with some periodic modulations, as one would expect. Our methodology allows us to freely change the **ξ** (Figure 8).

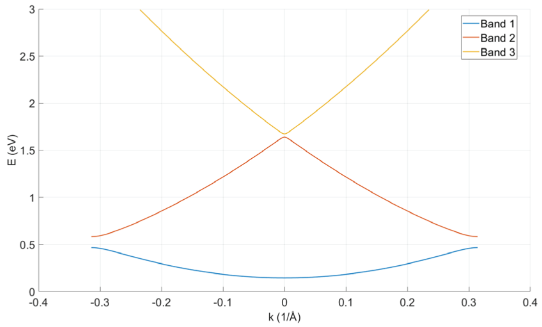


Figure 4: The band structure for **ξ = 0.3**, ***V_{top}* = 0.3eV**, and ***T* = 1nm**. One can see that the first band is the flattest. It will be the focus of the next analysis.

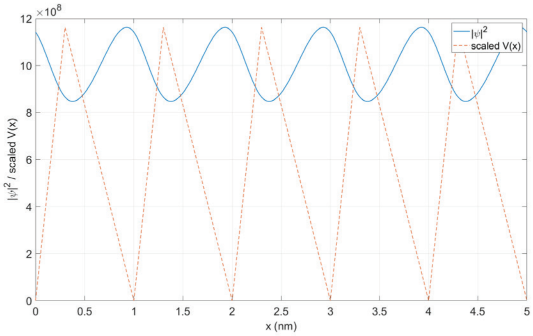


Figure 5: Probability density at E = 0.148eV for **ξ = 0.3**, ***V_{top}* = 0.3eV**, and ***T* =1nm**. The probability density has been shown in blue, and the potential is shown in orange for effect. Low-energy solutions are highly localized, as one would expect.

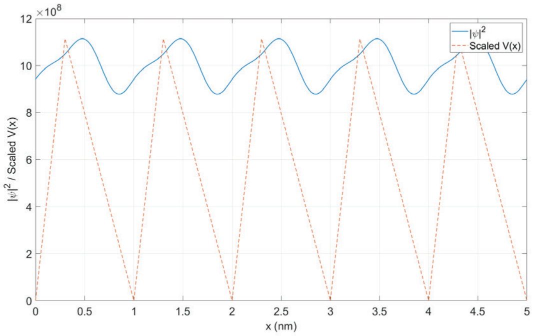


Figure 6: Probability density at E = 1.073eV for **ξ = 0.3**, ***V_{top}* = 0.3eV**, and ***T* =1nm**. As energy increases, the solutions are more reminiscent of plane waves.

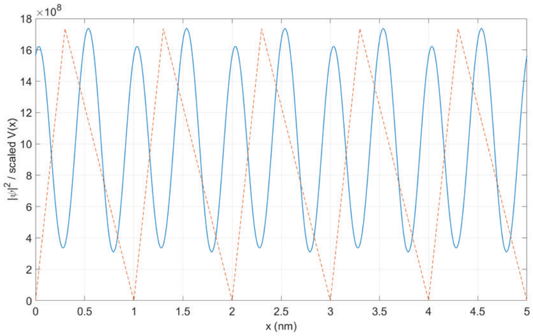


Figure 7: Probability density at E = 1.631eV for **ξ = 0.3**, ***V_{top}* = 0.3eV**, and ***T* =1nm**. At high energies, solutions resemble plane waves with periodic modulations.

DOI: 10.36838/v8i2.7

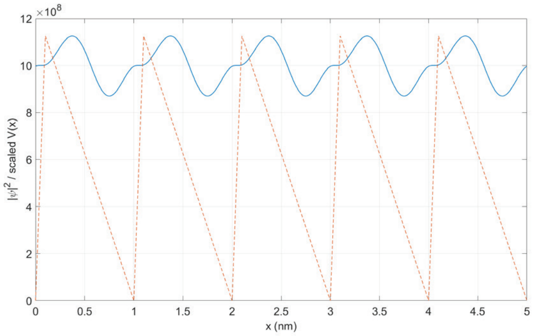


Figure 8: Probability density at E = 1.005eV for $\boldsymbol{\xi}$ **= 0.1**, $\boldsymbol{V_{top}}$ **= 0.3eV**, and $\boldsymbol{T}$ **=1nm**. Our methodology still works for more extreme values of $\boldsymbol{\xi}$.

Asymmetry and band structure:

We mentioned before that sawtooth potentials have been studied for their ability to create "flat bands", bands of very low bandwidth (slowly varying E with k) that are experimentally useful in studying weak interactions. By examining how the width of the lowest energy band varies with parameters of the lattice **(** $\boldsymbol{V_{top}}$, $\boldsymbol{\xi}$, and $\boldsymbol{T}$**)**, we better understand under what conditions flat bands are observed in sawtooth potentials.

We find no dependence of the bandwidth on $\boldsymbol{\xi}$ at a constant $\boldsymbol{V_{top}}$ **= 0.3eV**, and $\boldsymbol{T}$**=1nm** (Figure 8). This is quite an interesting result, but we can motivate it through a semi-classical treatment of the potential. The tunneling amplitude through a singular sawtooth, and thus bandwidth, is dependent on the integral of the square root of V(x) over the classically forbidden region for a singular sawtooth.[12,13] The classically forbidden region for a particle with $\boldsymbol{E < V_{top}}$ is given by $\boldsymbol{V_{top}/k_2 \le x \le E/k_2}$ and $\boldsymbol{E/k_1 \le x \le V_{top}/k_1}$. The WKB action over the region is then given by

$$S_L = \int_{V_{\text{top}}/k_2}^{E/k_2} \sqrt{2m\,(k_2\,x - E)}\,dx = -\frac{2\sqrt{2m}}{3k_2}\,(V_{\text{top}} - E)^{3/2}, \tag{27}$$

$$S_R = \int_{E/k_1}^{V_{\text{top}}/k_1} \sqrt{2m\,(k_1\,x - E)}\,dx = \frac{2\sqrt{2m}}{3\,k_1}\,(V_{\text{top}} - E)^{3/2}. \tag{28}$$

$$S = S_L + S_R = \frac{2\sqrt{2m}}{3}(V_{\text{top}} - E)^{3/2}\Big(\frac{1}{k_1} - \frac{1}{k_2}\Big) \tag{29}$$

Since $\boldsymbol{T = \frac{V_{top}}{k_1} - \frac{V_{top}}{k_2}}$, without changing T, the WKB action and thus the tunneling amplitude are not dependent on the asymmetry. Thus, we might expect the bandwidth of the first band not to depend on asymmetry $\boldsymbol{\xi}$. This is unique to the sawtooth potential: a general skewing of a potential function that leaves the area under the function unchanged would not leave the WKB action unchanged due to the square root in the WKB integral. When altering $\boldsymbol{\xi}$, one side becomes steeper but shorter, and the other becomes less steep but longer. The 2 effects happen to compensate perfectly in the case of linear functions so that the overall "tunneling difficulty" is unchanged.

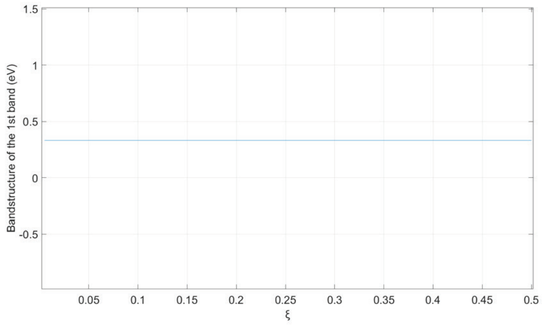


Figure 9: Bandwidth of first band (eV) vs $\boldsymbol{\xi}$ for $\boldsymbol{V_{top}}$ **= 0.3eV**, and $\boldsymbol{T}$ **= 1nm**. Bandwidth is not affected by $\boldsymbol{\xi}$ as rationalized through a semiclassical approach.

$\boldsymbol{V_{top}}$, T, and band structure:

The effect of potential height $\boldsymbol{V_{top}}$ and cell length $\boldsymbol{T}$ on band structure for a general periodic potential is well documented.[13] We will briefly present the effects of these parameters as applied to the sawtooth potential. We find that there is a strong correlation between the $\boldsymbol{V_{top}}$ and the bandwidth of the 1st band. As $\boldsymbol{V_{top}}$ increases, the tunneling amplitude across the unit cell decreases, and the bandwidth of the first band decreases (Figure 9). Similarly, as $\boldsymbol{T}$ increases, the tunneling amplitude decreases, thus decreasing the bandwidth of the 1st band (Figure 10).

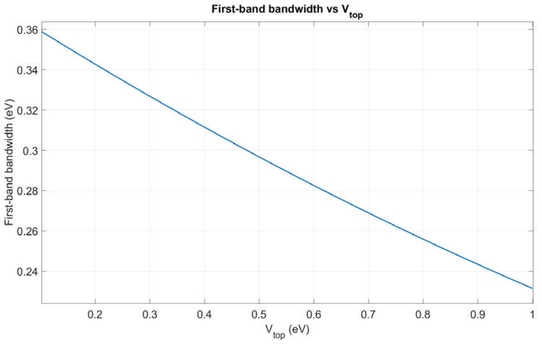


Figure 10: Bandwidth of first band (eV) vs $\boldsymbol{V_{top}}$ (eV) for $\boldsymbol{\xi}$ **= 0.3**, $\boldsymbol{T}$ **= 1nm**. Bandwidth decreases with $\boldsymbol{V_{top}}$, as one would expect.

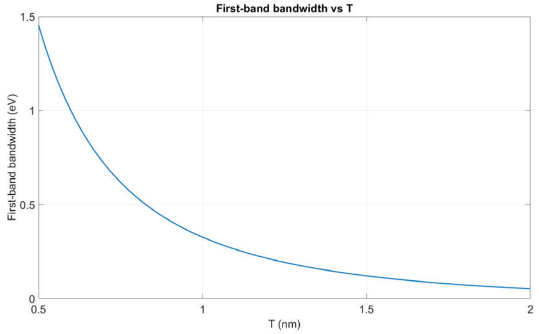


Figure 11: Bandwidth of first band (eV) vs $\boldsymbol{T}$ (nm) for $\boldsymbol{V_{top}}$ **= 0.3eV**, $\boldsymbol{T}$ **= 1nm**. Bandwidth decreases with $\boldsymbol{T}$, as one would expect.

■ Conclusion

The solutions to the time-independent Schrodinger equation for a sawtooth potential were presented and visualized. We applied the solutions and the band structure to investigate the relationship between the bandwidth of the first band and parameters V_{top} , ξ, and T. As is true for general periodic potentials, we found that increasing V_{top} or T decreased the bandwidth, but interestingly, that ξ had no impact on the bandwidth, which we then motivated with a semiclassical treatment of the system. These findings could offer insights into parameter choices for flat bands in sawtooth potentials, demonstrating the utility of the obtained solutions. Furthermore, being able to alter ξ without affecting the bandwidth would allow us to alter the transport properties of the sawtooth whilst retaining a flat band and amplifying weak interactions.[3-5] Next steps could include solving numerically using Airy functions instead of using a piecewise constant approximation to provide more accurate solutions. Further research could be conducted on irregular sawtooth potentials with individual sawtooth potentials of varying shapes and sizes.

■ Acknowledgments

Thanks to my mentor, Adam Almakroudi, for mentoring me throughout the research process and teaching me much of the pre-requisite knowledge required.

■ References

1. Ait-Haddou, R.; Herzog, W. Brownian Ratchet Models of Molecular Motors. Cell Biochemistry and Biophysics 2003, 38 (2), 191–214. https://doi.org/10.1385/cbb:38:2:191.
2. Hwang, W.; Karplus, M. Structural Basis for Power Stroke vs. Brownian Ratchet Mechanisms of Motor Proteins. Proceedings of the National Academy of Sciences of the United States of America 2019, 116 (40), 19777–19785. https://doi.org/10.1073/pnas.1818589116.
3. Salger, T.; Kling, S.; Hecking, T.; Geckeler, C.; Morales-Molina, L.; Weitz, M. Directed Transport of Atoms in a Hamiltonian Quantum Ratchet. *Science (New York, N.Y.)* 2009, 326 (5957), 1241–1243. https://doi.org/10.1126/science.1179546.
4. Zhang, T.; Jo, G.-B. One-Dimensional Sawtooth and Zigzag Lattices for Ultracold Atoms. Scientific Reports 2015, 5 (1). https://doi.org/10.1038/srep16044.
5. Pyykkönen, V.; Peotta, S.; Fabritius, P.; Mohan, J.; Esslinger, T.; Törmä; Törmä, P. ETH Library Flat-Band Transport and Josephson Effect through a Finite-Size Sawtooth Lattice Flat Band Transport and Josephson Effect through a Finite-Size Sawtooth Lattice. *Physical Review B 103* (14)
6. Hodžić, M.; Ulrich, M.; Graz, H. Numerical Simulation of Directed Quantum Tunnelling; 2021. https://static.uni-graz.at/fileadmin/_Persoenliche_Webseite/hohenester_ulrich/diplomahodzic21.pdf
7. Rozenbaum, V. M.; Shapochkina, I. V.; Trakhtenberg, L. I. Quantum Particle in a V-Shaped Well of Arbitrary Asymmetry. Brownian Motors. Physics-Uspekhi 2024, 67 (10), 1046–1055. https://doi.org/10.3367/ufne.2024.06.039704.
8. Vallée, O., & Soares, M. (2004). Airy Functions and Applications to Physics. IMPERIAL COLLEGE PRESS. https://doi.org/10.1142/p345
9. Kittel, C. Introduction to Solid State Physics, 8th ed.; Wiley: Hoboken, NJ, 2004; Chapter 9, "Energy Bands in the Nearly-Free Electron Model," Section 9.1, "Bloch Waves."
10. Strang, G. *Introduction to Linear Algebra*, 5th ed.; Wellesley–Cambridge Press: Wellesley, MA, 2016; Chapter 3, "Solving Ax = 0.
11. Almeida, J.; Rodrigues, T.; Bruno-Alfonso, A. Solving the Schrödinger Equation by the Transfer-Matrix Method. *Matemática Contemporânea* 2024, 59 (8). https://doi.org/10.21711/231766362024/rmc598.
12. Griffiths, D. J. *Introduction to Quantum Mechanics*, 3rd ed.; Cambridge University Press: Cambridge, UK, 2018; Chapter 11.
13. *Band structure, David Tong: Solid State Physics*. www.damtp.cam.ac.uk. https://www.damtp.cam.ac.uk/user/tong/solidstate.html.

■ Author

Zhixuan (Jason) Tao is an aspiring physics major based in New Zealand. Jason is interested in modern physics and, in particular, quantum mechanics, and looks forward to studying these concepts in college.

IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Retinal Photographs Improve the Diagnosis of Autism Spectrum Disorder

Seyoung Park
The Webb School, 1175 West Baseline Road, Claremont, CA, 91711, USA

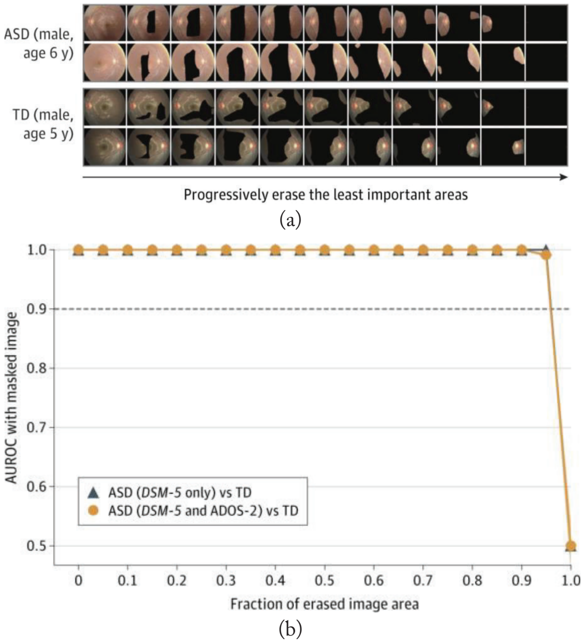
ABSTRACT: Autism spectrum disorder (ASD) is a neurological and developmental condition that affects behavior, communication, and learning, often requiring lifelong support. Diagnosis by the age of two years can significantly reduce symptom severity and enhance cognitive, language, and social skills. However, current diagnostic methods rely heavily on subjective behavioral observations, rendering them prone to inaccuracies, stressful for caregivers, and time-consuming. To address this issue, this study introduces a novel and objective diagnostic system that utilizes retinal (fundus) photographs in conjunction with machine learning. The fast gradient sign method (FGSM), originally developed as an adversarial perturbation technique, was applied in this study to evaluate the robustness of convolutional neural networks in classifying ASD from retinal images. This robustness test also resulted in modest performance improvements across all tested models, surpassing baseline performances. These findings could aid the development of efficient, accurate, and non-invasive tools for early ASD detection and intervention, thereby significantly benefiting individuals with ASD and their families. Future studies should investigate additional adversarial methods and incorporate larger and more diverse datasets.

KEYWORDS: Behavioral and Social Sciences, Neuroscience, Autism Spectrum Disorder, Retinal Photographs, Fast Gradient Sign Method.

Introduction

Autism spectrum disorder (ASD) is a neurological and developmental condition that affects behavior, learning, and communication. It comprises a wide variety of types and severities among patients. ASD is a lifelong disorder that requires ongoing management, although medication and treatments can lessen its severity.[1] Affecting one in 36 children,[2] ASD occurs across all ages, sexes, and ethnicities, rendering early screening highly recommended. Early interventions for ASD, ideally at the age of two or younger, can significantly improve cognitive, language, and social interaction abilities.[3]

Traditional ASD diagnosis is a two-step process that involves healthcare providers, caregivers, and children. Wellness checkups and visits to healthcare providers help identify symptoms of ASD. Children with abnormal birth conditions or a family history of ASD undergo more thorough screening. Although ASD diagnosis is considered an accurate process, it involves a long-term examination, assessment, and conversations that can distress both caregivers and children. As it is complex, these subjective evaluations are not suitable for all cases. To address this problem, a previous study examined the use of machine learning models with retinal or fundus photographs to screen for ASD and evaluate symptom severity, demonstrating a correlation between optic disc features and ASD diagnosis.[4] As shown in Figure 1 Kim *et al*. evaluated the model performance by progressively removing 5% of the fundus photographs that were considered the least important to observe the change in the area under the receiver operating characteristic curve (AUC-ROC). Surprisingly, even when 95% of the images were removed, no significant change was observed in the AUC -ROC. However, when the area with the optic disc was masked, the AUC -ROC decreased abruptly.



(a)

(b)

Figure 1: Previous studies on screening for ASD using fundus photographs. (a) Example of the data collection process; (b) Graph representing the AUC -ROC as the masked area of the image increases.[4]

Kim *et al* . demonstrated the feasibility of using machine learning and fundus photographs, particularly of the optic disc, to diagnose ASD. Nevertheless, their study had limitations, as the model was evaluated using only 1,890 eyes from 958 participants, which could be considered a relatively small dataset.

DOI: 10.36838/v8i2.13

Additionally, the human bias and subjectivity introduced through the use of handcrafted features demonstrated the inconsistency in conducting this experiment. Recent studies have also emphasized both the opportunities and challenges of early ASD screening. For example, Okoye *et al.* reviewed the clinical benefits and risks of early ASD diagnosis, while Kim *et al.* demonstrated the feasibility of applying deep learning to retinal images.[3,4] Furthermore, disparities in ASD diagnosis related to race and socioeconomic status have been documented, underscoring the need for objective and widely applicable diagnostic methods.[5-18] The present study provides a broader effort to improve accuracy, equity, and efficiency in ASD detection. In particular, the study developed a machine learning and mathematics-based adversarial technology that effectively used medical information. It focused on applying fundus photographs, particularly those of the optic discs, in systematic and mathematical approaches for ASD diagnosis.

ASD:

Individuals with ASD often experience difficult expressing themselves verbally and may rely on nonverbal body language.[1] Due to its detrimental effect on humans, ASD screening at a young age and early diagnosis are crucial in preventing severe impairment. Despite this need, the average age of ASD diagnosis in the United States is five years, even though ASD can be reliably diagnosed by specialists at age two.[5]

Traditional ASD diagnosis involves a thorough evaluation process, which includes collecting a developmental history from parents or caregivers, observing the child's behavior, and using standardized screening tools such as the Modified Checklist for Autism in Toddlers (M -CHAT). Professionals apply the DSM-5 criteria and administer assessments, such as the Autism Diagnostic Observation Schedule and the Autism Diagnostic Interview (ADI) . A team of professionals conducts an evaluation comprising parental interviews, developmental testing, and, if necessary, hearing tests, vision screening, and genetic testing. Throughout the process, continuous monitoring is provided to refine and adjust as needed.[6] Hence, diagnosing ASD is a longlong-term process that can be exhaust ing for both caregivers and children. Additionally, the assessment is subjective and does not guarantee complete accuracy.

Figure 2: Three scenes portrayportraying ASD diagnoses. (a) Hearing test for ASD screening[7]; (b) ASD screening with a therapist[8]; (c) A toddler undergoing ASD screening.[9]

A previous study demonstrated the potential of using fundus photographs for accurate and objective diagnosis of ASD severity. The calculated AUC-ROC values were 1.00 with a 95% CI for ASD screening and 0.74 with a 95% CI for symptom severity, indica ting that the model was highly reliable.[4] These results demonstrate the importance of accessible, time-efficient, and objective ASD screening and diagnosis.

Fundus Photographs:

Fundus photographs, also known as retinal photographs, show the fundus located at the back of the human eye. It comprises the retina, macula, fovea, optic nerve, and optic disc (Figure 3a). Fundus photography is easily performed in ophthalmology institutes using a fundus camera, which is a non -invasive, painless device. Colored fundus images are obtained and examined to determine the presence of diseases and disorders.[10] A recent device, depicted in Figure 3c, demonstrates a method for taking fundus photographs at home using a cell phone. These new devices, which have made fundus photography more accessible, and machine learning technology, offer a non-invasive approach for observation and diagnosis at a low cost in a flexible environment.

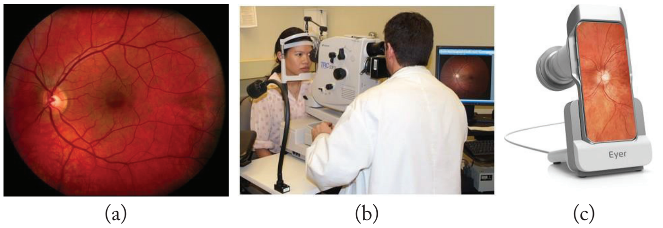


Figure 3: Fundus photograph. (a) Example of a colored fundus photograph[11]; (b) Traditional method of observing the eye in an ophthalmology institute[12]; (c) Mobile phone-based fundus imaging device.[13]

Adversarial Perturbation:

Adversarial perturbation is a crucial concept in machine learning, originally developed to test the robustness and vulnerability of neural networks by introducing small, imperceptible noise to the input data. Such perturbations, though invisible to the human eye, can lead to significant misclassifications, thereby exposing the limitations of neural network models.[14] Rather than serving as a traditional augmentation technique, adversarial perturbation is designed to challenge models under controlled distortions, enabling the evaluation of model stability and generalization. Among various adversarial methods, this study focused on the fast gradient sign method (FGSM), applying it as a robustness-oriented experiment to assess how convolutional neural networks respond to perturbations in retinal images.[15]

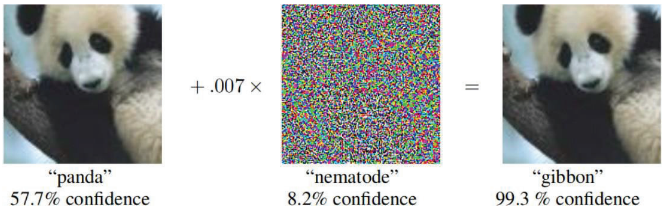


Figure 4: Adversarial perturbation.[1] Minimal adversarial noise (0.007 magnitudes) significantly shifts the model prediction from panda (57.7%) to gibbon (99.3%) illustrating the vulnerability of convolutional neural networks (CNN).

The FGSM utilizes the gradients of the loss function with respect to the input data to determine the perturbations applied to the input. For instance, the neural network identifies the image as that of a panda with a confidence level of 57.7% (Figure 4) . However, when a small amount of noise (denoted as 0.007 times a specific color pattern) is added by calculating the gradients of the loss function of the input image, it can effectively mislead the model into classifying the image ina ccurately. The model identifies the perturbation pattern as a nematode with 8.2% confidence, which is irrelevant, but illustrates the additional randomness of the adversarial pattern. Although the resulting image appears to be a panda to human eyes, it is classified as a gibbon with 99.3% confidence, indicating a misclassification by the neural network and a significant shift caused by a slight alteration in the input image.[16]

Methods

This study applied FGSM, a commonly used adversarial perturbation method, to add noise that disturbs the learning process. Three novel methods were explored: additivity of the FGSM attack on the fundus photograph, additivity of the FGSM attack solely on the optic nerve head of the photograph, and complete removal of the optic nerve head from the photograph. These methods were examined to estimate changes in the accuracy of the method when adversarial perturbation was added , as well as the role of the optic nerve head in ASD diagnosis.

Baseline:

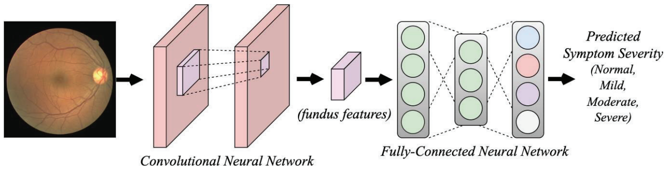


Figure 5: Baseline convolutional neural network (CNN) architecture for ASD severity classification.

Figure 5 illustrates the basic process for predicting symptom severity using the baseline model. This is the basic architecture of the classification network used in this study. The network uses a fundus photograph as an input $I \in RHW$ and generates feature maps. H and W denote the height and width of the fundus photograph, respectively. Fundus features, the output of a convolutional neural network, are represented as a three-dimensional matrix denoted by z € ⟦R⟧ ^l. This leads to the FCNN, which outputs different probability values for each of the four possibilities: normal, mild, moderate, and severe. As illustrated by the different colors in Figure 5, each element has a score value, which can be altered into a probability using the Softmax function. This probability represents the model's prediction of the possibility of this symptom range. This process can be defined as: $FCNN: Z \rightarrow P$.

Equation 1. Softmax function.

$$P_k = \frac{e^{S_k}}{\sum_j e^{S_j}}$$

Equation 1 illustrates the Softmax function, which converts a set of raw scores into probabilities that are easier to interpret and work with when utilizing machine learning. Pk is the output or the probability assigned to class k, Sk is the score for k, and $\sum_j e^{s^j}$ *It* is the sum of the exponentials of all the raw scores. By exponentiating each score, the equation checks all outputs. Normalizing these values by dividing by the sum of all the exponentials of the scores ensures that output probabilities, when added up, equal one.

Equation 2. Cross-entropy loss function.

$$L_{ce} = -log_e P$$

Equation 2 presents the cross-entropy loss function, which evaluates a model's performance by comparing its predicted probability distribution with the actual distribution. L_{ce} is the output of the cross-entropy loss or the probability value between 0 and 1, where log_e represents the natural logarithm, and P is the predicted probability of the correct class. Specifically, the loss value is quantified by taking the negative logarithm of the predicted value, thereby minimizing this loss value and improving the model's ability to make accurate predictions. A loss value closer to one indicates a lower loss, whereas a loss value closer to zero indicates a higher loss.

Proposed Noise Model (Fundus):

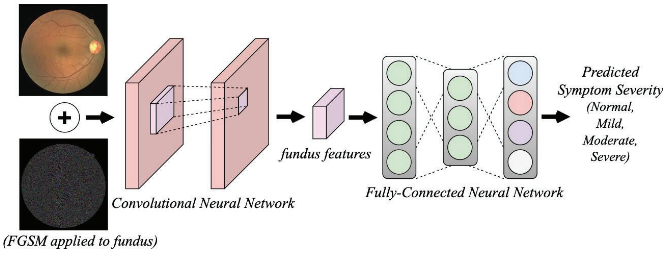


Figure 6: Noise model with fast gradient sign method (FGSM) applied to entire fundus images.

Figure 6 illustrates the architecture of the first additive proposed in this study for classifying ASD symptom severity. All processes were identical to the baseline architecture, except for the input of the image, which included an FGSM-applied fundus photograph. The FGSM is a picture comprising small dots of color, which makes no difference in how a human views the photo; however, it renders machine learning more challenging for computers. The noise value was denoted as N_{total}. This FGSM attack is mathematically constructed by reverse-engineering a typical gradient-descent algorithm. A typical gradient descent algorithm iteratively uses input and gradient values to produce a better optimized result through extensive calculations. Instead of using the gradient descent algorithm to increase our output value positively, the gradient values are included in N_{total} to make training more difficult. To improve the results, the noise value increases in every sample.

Proposed Noise Model (Optic Nerve Head):

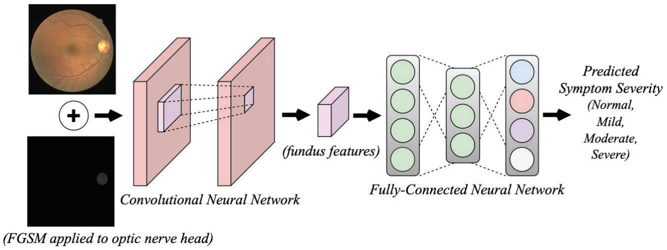

+
Convolutional Neural Network
(fundus features)
Fully-Connected Neural Network
Predicted
Symptom Severity
(Normal,
Mild,
Moderate,
Severe)
(FGSM applied to optic nerve head)

Figure 7: Noise model with FGSM applied only to the optic disc region.

Figure 7 shows the architecture of the second additive proposed in this study, in which noise was added to the optic nerve head area. Similar to the first additivity, all processes are identical except for the input, which is a picture with noise or FGSM applied solely to the optic disc of the fundus photograph. The input is denoted by N_{disc}. As the optic disc is a crucial part of ASD diagnosis, it can be hypothesized that the accuracy does not increase by adding noise to this specific system. However, this experiment further investigated whether the accuracy would be maintained by adding noise to the optic disc specifically, rather than the entire fundus photograph, and, if not, the rate of decrease in comparison to the proposed noise model in Equation 2. This experiment also used a cross-entropy loss function.

Proposed Noise Model (Optic Nerve Head Removed):

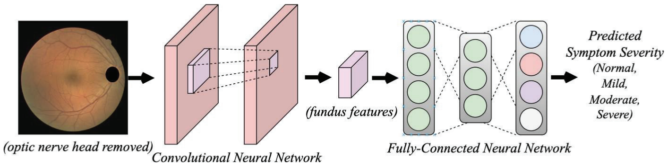

(optic nerve head removed)
Convolutional Neural Network
(fundus features)
Fully-Connected Neural Network
Predicted
Symptom Severity
(Normal,
Mild,
Moderate,
Severe)

Figure 8: Noise model with the optic disc region removed from fundus images.

Figure 8 illustrates the architecture of the last model, in which the optic nerve head has been removed from the photograph. The process is identical; however, the input differs because the optic nerve head is completely removed. This input is referred to as I_{nodisc}. It is created through simple coding by changing the existing pixel values of the optic disc to black, given the disc area.

Fundus Dataset:

This study used a dataset from the AI Hub, a government-funded database in Korea, to conduct experiments.[17] The most recently updated version of the data dated January 19, 2024 was utilized. From 1,038,674 samples representing various diagnoses of disorders in children and adolescents, 57,195 samples consisting solely of fundus photographs from children and adolescents with ASD and those without any diagnosed disorders were collected.

Although the samples were exclusively from South Korea and comprised data on South Koreans, this should not affect the accuracy of the experiment, as ASD is not correlated with a specific ethnicity.[18] Among the 57,195 samples, 37,145 (64.9%) were normal and consisted of fundus photographs of children and adolescents without the disorder, and 20,050 (35.1%) were samples of children and adolescents with ASD. In terms of age distribution, 27.70% of the samples were from children under seven years of age, 43.57% were from children aged 7–12 years, and 28.73% were from adolescents aged 13–20 years. Additionally, the ratio of the samples used for training and testing was 8:2.

Result and Discussion

Evaluation of the FGSM Applied Model:

Table 1: Evaluation result of the FGSM applied model (fundus): a comparison of performance metrics (accuracy, recall, precision, and F1-score) across four CNN architectures under adversarial perturbation applied to entire fundus images.

FGSM (Fundus)	Accuracy	Recall	Precision	F1-Score
ConvNeXt[19]	86.70	86.12	85.97	86.04
DenseNet-201[20]	87.55	87.12	86.60	86.86
ResNet-101[21]	87.93	87.34	86.55	86.94
ResNet-152[21]	89.11	88.58	87.82	88.20

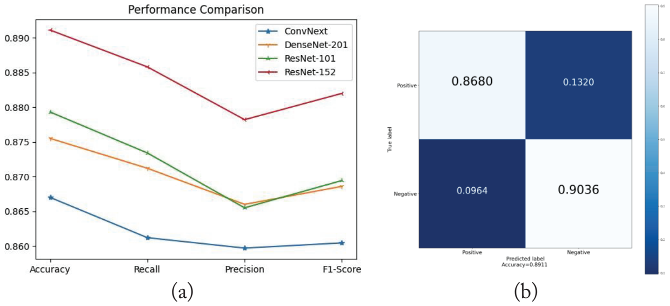


Figure 9: Evaluation results. (a) Comparison of accuracy, recall, precision, and F1-score across four CNN architectures with FGSM applied to the entire fundus image, showing that ResNet-152 outperforms the others in all metrics. (b) Confusion matrix for ResNet-152 under FGSM perturbation, achieving 89.11% accuracy. High true-positive and true-negative rates indicate strong robustness to adversarial noise.

Figures 9a and 9b, along with Table 1, present the performance comparison graph, confusion matrix, and summary table, respectively, offering insights into the experimental results of evaluating the performance of various CNNs, particularly focusing on how they handle adversarial perturbations (FGSM) in the symptom severity assessment of ASD through fundus photographs. Figure 9a shows the performance metrics (accuracy, recall, precision, and F1-score) for the four CNN architectures: ConvNeXt, DenseNet-201, ResNet-101, and ResNet-152. ResNet-152, which is a deep network with 152 layers, outperformed the other models across all metrics, indicating that it was the most effective model for this task.

ConvNeXt, with 50 layers, showed the lowest performance, particularly in terms of recall (a measure of true positives from all positive samples) and precision (a measure of positive predictions), suggesting that it may not be as reliable for correctly identifying both positive and negative cases. The table in Figure 9a provides a detailed breakdown of the performance comparison graph for each model, reinforcing the observation that ConvNeXt lags behind the other models, presumably because of its shallow layers, which hinder complex studies with FGSM.

DOI: 10.36838/v8i2.13

Thus, the findings suggest that machine learning models, particularly deep-learning CNNs, can serve as powerful tools for ASD screening and symptom severity assessment by accurately analyzing retinal images. Spesifically, the confusion matrix shown in Figure 9b indicates an overall accuracy of 89.11%. The gradation scale measures accuracy, with dark blue indicating the least accurate and light blue indicating the most accurate. Visually, the true-positive (correctly identified ASD patients using fundus photographs) and true-negative (correctly identified non-ASD patients using fundus photographs) rates were low. In contrast, there were false-positive results (identified as non-ASD patients with ASD). The false-negative rates (identifying ASD patients as non-ASD) are low, indicating that the model correctly identifies the majority of the given dataset.

Therefore, the application of FGSM should primarily be interpreted as a robustness test, assessing the stability of CNN models under perturbation. The observed improvements in accuracy suggest that, beyond withstanding adversarial noise, the models demonstrated enhanced generalization. This reframing highlights FGSM's role in testing model robustness rather than serving as a conventional data augmentation method. These results confirm that deep-learning CNNs, particularly ResNet-152 among the ones tested, are effective tools for ASD screening and symptom assessment.

Optic Disc Area Applied:

Table 2: Comparison of model accuracy when the FGSM perturbation is applied only to the optic disc area versus baseline without perturbation. All four CNN architectures showed accurate improvements, with ResNet-152 achieving the highest increase (3.06%). These results suggest that localized perturbation to the optic disc can enhance model performance.

FGSM (Optic disc)	FGSM (Optic Disc; Accuracy)	Baseline (Accuracy)
ConvNeXt	86.5	83.54
DenseNet-201	87.54	84.18
ResNet-101	87.82	84.34
ResNet-152	88.89	85.83

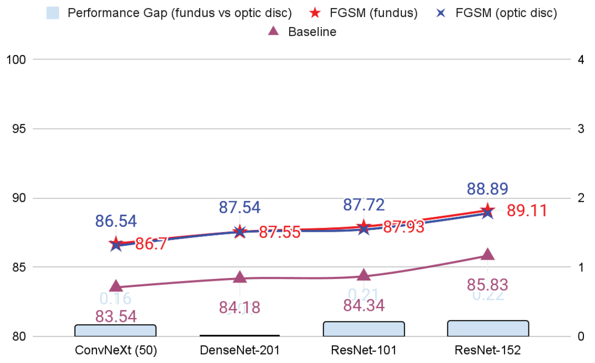


Figure 10: Ablation study results. Accuracy comparison across four CNN architectures for baseline, FGSM applied to the entire fundus, and FGSM applied only to the optic disc. Both FGSM conditions improved accuracy over baseline, with full-fundus FGSM showing slightly higher gains. ResNet-152 achieved the highest accuracy in all settings, indicating strong robustness to perturbation.

This ablation study compared the effects of applying FGSM solely to the optic disc area with a baseline (without FGSM) model, using a different approach to define the impact of the optic disc on diagnosing ASD symptom severity. It also compares the accuracies of four different neural networks in this architecture. Applying the FGSM to the optic disc resulted in performance improvements for all tested models. Specifically, ConvNeXt, which was the least accurate when tested in baseline architecture, showed a performance increase of 2.96, whereas ResNet-152 maintained its high accuracy with an increase of 3.06. These improvements highlight the efficacy of the FGSM in enhancing model performance by effectively preprocessing the input image. The graph further emphasizes the performance gap when the FGSM was added to the baseline model, highlighting that applying the FGSM to the fundus photograph was more accurate than applying it solely to the optic disc. However, for DenseNet-201, the performance gap was 0.01, indicating that applying the FGSM to optic discs or full fundus photography did not make a noticeable difference in diagnosing the severity of ASD symptoms.

Evaluation of Optic Nerve Head Removal Model:

Table 3: Accuracy comparison across four CNN architectures for baseline and when the optic nerve head is entirely removed from fundus images. The removal of the optic disc resulted in a substantial performance drop for all models, with ConvNeXt showing the most significant decrease (–9.07%) and ResNet-101 showing the smallest (–6.67%). These results highlight the critical role of optic disc information in ASD severity classification from retinal images.

	Optic Nerve Head Removed (Accuracy)	Baseline (Accuracy)
ConvNeXt	74.47	83.54
DenseNet-201	76.35	84.18
ResNet-101	77.67	84.34
ResNet-152	77.98	85.83

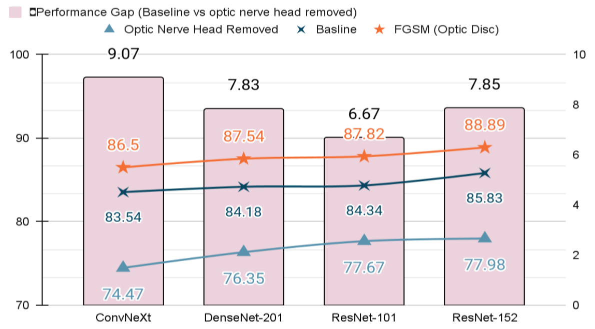


Figure 11: Evaluation result of optic nerve head removal experiment: accuracy comparison across CNN architectures, highlighting the performance drop when the optic nerve head is removed versus the baseline and FGSM (optic disc) conditions.

The second ablation study focused on the performance of various neural network architectures when the optic nerve head was removed entirely from the fundus photographs. As shown in Table 3, the accuracy of the different models in diagnosing the symptom severity of ASD decreased drastically to 70% when the optic nerve head was removed from the input image, compared to the baseline (the original study without FGSM applied). ConvNeXt, with 50 layers, showed the most significant performance drop of 9.07, indicating a high dependency on optic nerve head information. Simultaneously, ResNet-101

DOI: 10.36838/v8i2.13

experienced the smallest performance drop of 6.67, suggesting greater robustness in removing this feature. Overall, all neural network architectures experienced a significant decline in performance, underscoring the importance of the optic nerve head in medical imaging tasks, particularly in the diagnosis of ASD.

Figure 11 summarizes the performance gap between the baseline, optic nerve head removed, and FGSM applied to the input images. The FGSM results consistently showed a higher performance than both the baseline and the removed optic nerve head, suggesting that FGSM may be a more effective preprocessing method for enhancing model performance. Among the various models, ResNet-152 demonstrated the highest accuracy in evaluating fundus photographs for both the baseline and FGSM, underscoring the significance of the depth of the neural network in its performance. These insights provide a step-ahead solution for accurately diagnosing symptom severity in ASD, which is valuable for future model design and selection in medical imaging applications.

■ Conclusion

This study proposed and evaluated a novel system to diagnose ASD symptom severity using fundus photographs, focusing on the optic disc and the FGSM. This study applied FGSM as a robust-oriented perturbation technique to the entire fundus photograph and specifically to the optic disc, demonstrating that adversarial perturbation enhanced smodel performance. Furthermore, this study systematically the significance of the optic disc by comparing the accuracy of ASD diagnosis following its complete removal. The findings revealed that applying the FGSM to the optic disc significantly improve diagnostic accuracy across multiple neural network architectures, surpassing the baseline performance. Performance noticeably declined when the optic disc was removed entirely, underscoring the critical role of the optic disc in medical imaging tasks. Moreover, an analysis of various models showed that the deeper layers of feature maps were correlated with performance accuracy. Overall, findings could help develop robust, effective, and non-invasive diagnostic tools for ASD, thereby improving early detection and intervention strategies.

Despite these promising results, this study has several limitations. First, the dataset was limited to pediatric and adolescent fundus images from South Korea. Further validation on more diverse, multi-ethnic cohorts is necessary. Second, while FGSM perturbations were useful as a robustness test, they represent only one type of adversarial approach; future research should explore additional techniques such as PGD (Projected Gradient Descent) or DeepFool. Finally, this study focused exclusively on retinal imaging. Future studies integrating multimodal data (e.g., genetic, behavioral, or linguistic features) could enhance diagnostic performance. Addressing these limitations will be essential to ensure the clinical applicability of this approach.

■ References

1. National Institute of Mental Health. *Autism Spectrum Disorder*. https://www.nimh.nih.gov/health/topics/autism-spectrum-disorders-asd
2. Centers for Disease Control and Prevention. *Data and Statistics on Autism Spectrum Disorder*. https://www.cdc.gov/autism/data-research/index.html
3. Okoye, C.; Obialo-Ibeawuchi, C. M.; Obajeun, O. A.; Sarwar, S.; Tawfik, C.; Waleed, M. S.; Wasim, A. U.; Mohamoud, I.; Afolayan, A. Y.; Mbaezue, R. N. Early diagnosis of autism spectrum disorder: a review and analysis of the risks and benefits. *Cureus* **2023**, 15 (8), e43226. DOI: 10.7759/cureus 43226
4. Kim, J. H.; Hong, J.; Choi, H.; Kang, H. G.; Yoon, S.; Hwang, J. Y.; Park, Y. R.; Cheon, K.-A. Development of deep ensembles to screen for autism and symptom severity using retinal photographs. *JAMA Network Open* **2023**, *6* (12), e2347692–e2347692. DOI: 10.1001/jamanetworkopen.2023.47692
5. Autism Speaks. *Autism Statistics and Facts*. https://www.autismspeaks.org/autism-statistics-asd
6. NYU Langone Health. *Diagnosing Autism Spectrum Disorder in Children*. https://nyulangone.org/conditions/autism-spectrum-disorder-in-children/diagnosis
7. Special Learning. *Hearing Evaluation for Children with Autism*. https://special-learning.com/hearing-evaluation-for-children-with-autism
8. Waypoint Wellness Center. *Autism Screenings*. https://www.waypointwellnesscenter.com/services/autism-screenings
9. Otto, F. Toddler screening essential for autism detection despite the national task force's reservations. *Drexel News*, February 22, 2016. https://drexel.edu/news/archive/2016/february/early_autism_screening_rec
10. Dhanorkar, A. What is the purpose of fundus photography? *MedicineNet*. www.medicinenet.com/what_is_the_purpose_of_fundus_photography/article.htm
11. Carver College of Medicine, Department of Ophthalmology and Visual Sciences. *Diagnostic Imaging Services*. https://eye.medicine.uiowa.edu/patient-care/imaging-services
12. The University of British Columbia. *Color Fundus Photography*. https://ophthalmology.med.ubc.ca/patient-care/ophthalmic-photography/color-fundus-photography
13. RO Staff. A 'smart' way to get into fundus photography. *Review of Optometry*, January 23, 2021. https://www.reviewofoptometry.com/article/jumpstart-healing-with-new-amniotic-membrane-graft-1
14. Nightfall. *Adversarial Attacks and Perturbations*. https://www.nightfall.ai/ai-security-101/adversarial-attacks-and-perturbations
15. Goodfellow, I. J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv* **2014**, *1412.6572*. DOI: 10.48550/arXiv.1412.6572
16. Knagg, O. Know your enemy: Why adversarial examples are more important than you realize. *Medium*, January 2, 2019. https://medium.com/data-science/know-your-enemy-the-fascinating-implications-of-adversarial-examples-5936bccb24af
17. AI Hub. *Fundus Image Data for Symptoms of Mental Illness in Children and Adolescents*. https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=71516
18. Durkin, M. S.; Maenner, M. J.; Baio, J.; Christensen, D.; Daniels, J.; Fitzgerald, R.; Imm, P.; Lee, L.-C.; Schieve, L. A.; Van Naarden Braun, K. Autism spectrum disorder among US children (2002–2010): socioeconomic, racial, and ethnic disparities. *American Journal of Public Health*, 2017, 107(11), 1818–1826. DOI: 10.2105/AJPH.2017.304032
19. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A Convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, June 18–24, 2022; IEEE: New York, NY, USA, 2022; pp 11976–11986. DOI: 10.1109/CVPR52688.2022.01167

20. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K. Q. Densely Connected Convolutional Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, July 21–26, 2017; IEEE: New York, NY, USA, 2017; pp 4700–4708. DOI: 10.1109/CVPR.2017.243
21. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, June 27–30, 2016; IEEE: New York, NY, USA, 2016; pp 770–778. DOI: 10.1109/CVPR.2016.90

■ Author

Seyoung Park attends The Webb School in Claremont, CA. She recently won first place at the California Science Fair (Behavior Medicine) for the research presented in this paper. She has previously published research on medical imaging classification using machine learning. She aspires to study science to impact human health positively.

IJHSR ijhighschoolresearch.org

RESEARCH ARTICLE

Healthcare Finance and Climate Risk

Albert C. Luo

The Webb Schools, 1175 W Baseline Rd, Claremont, California, 91711, USA; albertluo777@gmail.com

ABSTRACT: This paper examines the impact of climate risk on health expenditure. Using pooled regression analysis across multiple countries, the study finds that higher climate risk, measured through carbon dioxide emissions and other proxies, generally leads to a significant increase in health expenditure. The results suggest that climate-related adversity contributes to the rising medical expenses and strains the healthcare budget. This study recommends proactive investment in climate-resilient healthcare infrastructure and mitigating climate risk to prevent long-term costs and ensure sustainability in healthcare access.

KEYWORDS: Behavioral and Social Sciences, Healthcare Finance, Climate Risk.

Introduction

Nearly every individual worldwide has been directly or indirectly affected by the healthcare industry. Therefore, health expenditure is a critical component of public policy in how it is shaping the accessibility, quality, and sustainability of healthcare systems worldwide. For policymakers, understanding the factors driving healthcare costs is essential for ensuring the efficient allocation of resources and promoting public well-being. When examining existing literature, there is a lack of studies that use climate risk as a determinant of health expenditures, going beyond the traditionally examined variables such as economic growth, demographic shifts, technological advancements, and policy reforms.[1] Therefore, this paper fills the gap in the existing literature by focusing on climate risk as a determinant of health expenditure, using carbon emissions as one of the proxies (though carbon emissions are globally distributed and closely correlated with national income, this paper addresses these concerns by controlling for GDP per capita and other macroeconomic indicators).

Climate risk is a growing threat to public health and economic stability. Rising temperatures, extreme weather events, and environmental degradation contribute to the spread of infectious diseases, respiratory conditions, and heat-related illnesses, all of which impose substantial financial burdens on healthcare systems. Moreover, climate-related disasters exacerbate healthcare disparities by disproportionately affecting vulnerable populations, further worsening this already prominent issue. Given these implications, policymakers must integrate climate risk into healthcare planning to mitigate long-term negative effects and promote the well-being of the rest of the population.

The contribution of this paper is multifaceted. First, it extends the literature on health expenditure determinants by further exploring climate risk as a determinant, offering new insights into how environmental factors shape public health financing. Second, it provides policy recommendations for integrating climate resilience into healthcare budgeting, equipping policymakers with evidence-based strategies to mitigate climate-induced health costs. By bridging the gap between climate economics and health policy, this study underscores the need for interdisciplinary approaches to address emerging global challenges. In summary, this paper advances the study of climate risk and health expenditure by using multiple proxies and employing richer analytical techniques across multiple countries.

Literature Review

The studies reviewed span diverse regions, including OECD countries,[2,3] G7 countries,[4] African nations,[5-7] Asian countries,[8] and European countries.[9] Several also focus on specific nations such as Russia,[10] China,[11-13] Bangladesh,[14-16] Spain,[1] and the United States.[17,18] While most of these papers primarily investigate the determinants of health expenditure, none incorporated climate risk as a central variable.[11] This paper aims to address this gap by focusing on OECD countries to offer a comprehensive understanding of the factors shaping health expenditure. Existing literature on health expenditure typically explores determinants such as economic indicators,[1,5,19,20] demographic variables,[8] and health system characteristics.[1],[2] Economic factors—GDP, per capita income, and wage growth—are among the most commonly cited determinants.[1,5,19,20] Demographics and health system features have also been widely analyzed.[1,2,8] Studies that include climate variables, by contrast, tend to examine their influence on other indicators like education or household consumption rather than directly linking them to health expenditure.[11,14]

This pattern is reflected in the works of Gao *et al.*,[11] Islam *et al.*,[14] and Leppänen *et al.*[10] Gao *et al.*[11] assess the effect of climate risk on regional education spending in China, uncovering spatial dependencies and disparities in response across provinces. Islam *et al.*[14] explore how repeated climatic shocks influence household expenditures in Bangladesh, leading to significant reductions in food and non-food consumption. Leppänen *et al.*[10] evaluate how temperature fluctuations affect regional government spending in Russia, identifying reduced costs in colder regions and higher expenditures in warm areas. Though none of these studies directly address health expenditure, their insights into climate risk's broader socioeconomic implications

DOI: 10.36838/v8i2.20

highlight the relevance of further investigating health-related impacts in conjunction with climate risk.

Methodologically, the studies' approaches vary widely. Panel data analysis is used by Islam *et al*. and Dritsaki and Dritsaki to address unobserved heterogeneity across time and space.[4,14] Regression models are widely applied, including in studies by Bae *et al*.,[15] Chen *et al*.,[12] and Ampon-Wireko *et al*.[21] Hartwig and Sturm utilize Extreme Bounds Analysis (EBA) to test the robustness of economic determinants.[20] Gao et al. apply spatial econometric models to account for geographic dependencies.[11] Quantile regression, as used by Wang and Chen *et al*.,[12,19] provides insight into distributional effects across different levels of health expenditure. O'Neill *et al*.[22] take a distinct approach by using the Shared Socioeconomic Pathway Middle of the Road scenario (SSP2) to explore the link between educational attainment and climate resilience. Chaabouni and Saidi implement simultaneous equation models and GMM to examine causal interactions between CO_2 emissions, economic growth, and health spending.[23] The methodological diversity across studies offers a multifaceted perspective, revealing both strengths and limitations in assessing these complex relationships.

Despite differences in scope and method, the literature consistently identifies economic growth as a primary driver of health expenditure, with GDP and income levels emerging as robust predictors. However, the elasticity of this relationship differs from region to region. For example, in African nations, a 10% increase in GDP is associated with a 1% rise in health spending,[5] whereas studies from OECD countries suggest more elastic responses.[2] Demographic factors—especially aging—present mixed findings. For instance, while some suggest older populations elevate healthcare costs,[8] others emphasize the role of proximity to death and medical technology.[2] Additionally, many studies point out the importance of structural features such as governance models, insurance coverage, and fiscal autonomy in shaping national health expenditure.[1,20] What remains notably absent in this expansive literature is a direct exploration of how climate risk influences health expenditure.

Overall, the literature underscores the complex interaction between economic, demographic, and institutional factors in shaping health expenditure. However, the absence of studies directly connecting climate risk to health expenditure reveals a critical gap. This paper seeks to fill that gap by examining climate risk as a determinant of health expenditure within OECD countries, offering new insights into how environmental factors intersect with health system sustainability.

■ Methods

Model:

$$Y_i = \beta_0 + \beta_1 X1_i + \beta_2 X2_i + \beta_n Xn_i + \varepsilon_i$$

Y_i is the dependent variable for observation i, which refers to current health expenditure. β_0 is a constant term representing the expected value of a dependent variable when all independent variables are zero. β_1 to β_n are the coefficients for independent and control variables, which include climate risk as an independent variable and life expectancy at birth, inflation, GDP growth, government expenditure on education, real effective exchange rate index, age dependency ratio, out-of-pocket expenditure, and population as control variables. While keeping all other variables constant, each coefficient chooses how much the dependent variable changes when the corresponding independent variable changes by one unit. ε_i is the error term, which represents the difference between the actual value and the predicted value from the model.

■ Result and Discussion

Data:

Appendix 1 presents health expenditure, climate risk, and various economic indicators sourced from the World Bank World Development Indicators and Our World and Data.[24,25] The data sample covers the period from 1974 to 2022. The variable includes current health expenditure as a percent of GDP, climate risk as carbon dioxide emission and ND-GAIN, life expectancy at birth, inflation as an annual percentage of consumer prices, GDP growth, government expenditure on education as a percentage of total GDP, real effective exchange rate index, age dependency ratio as a percentage of working-age population, out-of-pocket expenditure as a percentage of current health expenditure, and population. Following existing literature, this paper estimates a pooled regression analysis to investigate the relationship between health expenditure and climate risk.

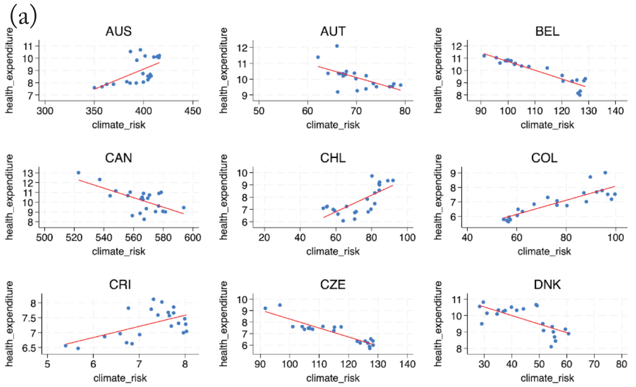
Findings:

Table 1 provides descriptive statistics for current health expenditure, climate risk, and other key indicators used in the analysis. Current health expenditure has 854 observations with a mean of 8.405 and a standard deviation of 2.207, ranging from a minimum of 3.855 to a maximum of 18.756. Climate risk, as annual total emissions of carbon dioxide, has 1862 observations with a mean of 325.19 and a standard deviation of 858.094, ranging from a minimum of 1.543 to a maximum of 6132.183. Inflation as an annual percentage of consumer prices has 1802 observations with a mean of 12.489 and a standard deviation of 55.082, ranging from a minimum of -4.448 to a maximum of 1281.443. GDP growth as an annual percentage has 1781 observations with a mean of 2.666 and a standard deviation of 3.618, ranging from a minimum of -32.119 to a maximum of 24.475. Government expenditure on education as a percentage of GDP has 1308 observations with a mean of 4.985 and a standard deviation of 1.229, ranging from a minimum of 0 to a maximum of 8.614. The real effective exchange rate index has 1446 observations with a mean of 99.37 and a standard deviation of 16.64, ranging from a minimum of 43.112 to a maximum of 194.383. The age dependency ratio as a percentage of the working-age population has 1900 observations with a mean of 52.981 and a standard deviation of 7.994, ranging from a minimum of 36.479 to a maximum of 99.671. Out-of-pocket expenditure as a percentage of current health expenditure has 842 observations with a mean of 20.744 and a standard deviation of 9.096, ranging from a minimum of 7.138 to a maximum of 55.664. The population in terms of people in a country has 1862 observations, with a mean of 30973870 and a standard deviation of 50209359, ranging from a minimum of 215291 to a maximum of 338000000.

Table 1: Numerical statistics for all variables examined in this research paper, including the number of observations, mean, standard deviation, minimum, and maximum values. The extensive data set reduces the source of errors in findings.

Variable	Obs	Mean	Std. Dev.	Min	Max
He	854	8.405	2.207	3.855	18.756
CO_2	1862	325.19	858.094	1.543	6132.183
Inf	1802	12.489	55.082	-4.448	1281.443
GDP Growth	1781	2.666	3.618	-32.119	24.475
Gov Exp Edu	1308	4.985	1.229	0	8.614
Exchange Rate	1446	99.37	16.64	43.112	194.383
Age Dep	1900	52.981	7.994	36.479	99.671
Out Of Pocket	842	20.744	9.096	7.138	55.664
Population	1862	30973870	50209359	215291	338000000

Figure 1 presents a positive correlation between climate risk and health expenditure across countries such as Australia (AUS), Chile (CHL), Colombia (COL), Costa Rica (CRI), Hungary (HUN), Luxembourg (LUX), Latvia (LVA), and New Zealand (NZL). This means that as climate risk increases in these countries, current health expenditure also increases. Possible explanations for the data range from extreme climate-related challenges to a lack of healthcare infrastructure or a mix of both. Negative Correlation is seen across countries such as Austria (AUT), Belgium (BEL), Canada (CAN), Czech Republic (CZE), Denmark (DNK), Germany (DEU), Estonia (EST), Finland (FIN), France (FRA), Ireland (IRL), Iceland (ISL), Israel (ISR), Italy (ITA), Japan (JPN), South Korea (KOR), Mexico (MEX), Norway (NOR), Poland (POL), Portugal (PRT), Slovakia (SVK), Slovenia (SVN), Sweden (SWE), Turkey (TUR), Great Britain (GBR), and the United States (USA). Possible explanations for the data range from well-developed healthcare systems built to handle health challenges attributed to climate risk to established climate adaptation strategies that mitigate the health impact of climate change. In general, more developed nations with higher GDPs, such as Germany (DEU), Canada (CAN), and the United States (USA), are expected to show a negative correlation. Smaller or more vulnerable countries, such as New Zealand (NZL) and Costa Rica (CRI), are expected to show a positive correlation–emphasizing their disproportional effects.



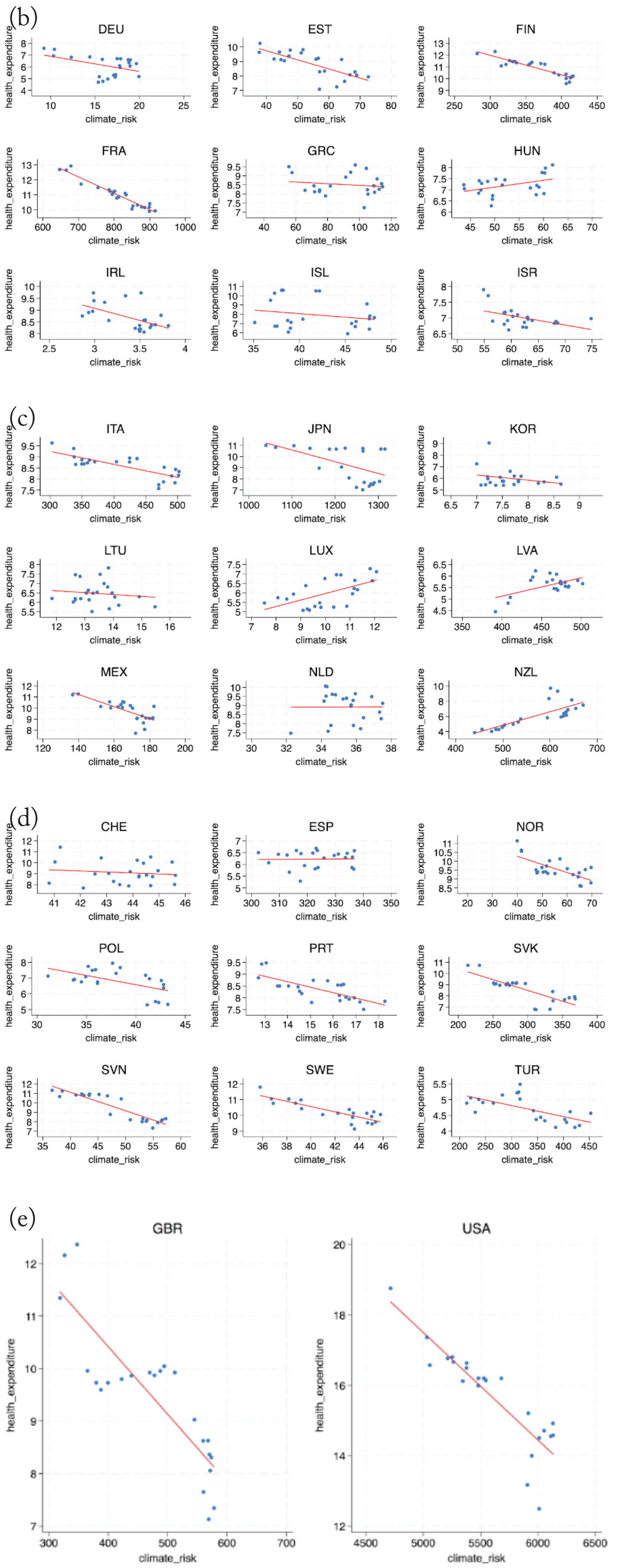


Figure 1: The scatter plot illustrates the relationship between current health expenditure and climate risk for 38 individual OECD countries: *a) AUS, AUT, BEL, CAN, CHL, COL, CRI, CZE, DNK; b) DEU, EST, FIN, FRA, GRC, HUN, IRL, ISL, ISR; c) ITA, JPN, KOR, LTU, LUX, LVA, MEX, NLD, NZL; d) CHE, ESP, NOR, POL, PRT, SVK, SVN, SWE, TUR; e) GBR, USA.* It finds a dynamic of relationships between different countries.

DOI: 10.36838/v8i2.20

Table 2 presents the results of four regression models. The dependent variable is the current health expenditure as a percentage of GDP (He), and the independent variable is CO2 as a proxy of climate risk. Each column represents different models with different control variables such as Inf, a proxy of economic stability; GDP growth, a proxy of economic development; and Government expenditure on education, a proxy of human capital development. In model 1, climate risk is positively associated with current health expenditure, with a coefficient of 0.0013–equivalent to an increase of approximately $0.40 per capita in OECD countries.

The result at the 1% level statistically emphasizes a high positive correlation between climate risk and current health expenditure, which suggests that higher climate risk leads to higher health expenditure. In model 2, after controlling for the effect of inflation, climate risk is still positively associated with current health expenditure at a 1% level. However, in several models, inflation shows a negative effect at the 1% level, indicating a negative correlation between inflation and health expenditure. In model 3, in addition to inflation, the effect of GDP growth on the effect of health expenditure is controlled. The result indicates a constant positive and statistically significant effect of climate risk on health expenditure. In this case, GDP growth negatively correlates with health expenditure at a 1% level. Finally, in model 4, with the additional control variable of government expenditure on education, the model continues to highlight the robust contribution of climate risk to health expenditure. Government expenditure on education also positively and significantly affects current health expenditure, with a p-value below 0.01.

Table 2: Four regression models analyzing current health expenditure show a strong positive correlation with climate risk. It provided evidence that a change in health expenditure is directly correlated with climate risk.

	(1)	(2)	(3)	(4)
VARIABLES	He	He	He	He
CO_2	0.0013***	0.0013***	0.0013***	0.0014***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Inf		-0.1497***	-0.1318***	-0.2236***
		(0.0264)	(0.0215)	(0.0207)
GDP Growth			-0.1496***	-0.1004***
			(0.0197)	(0.0183)
Gov Exp Edu				0.5777***
				(0.0391)
Constant	7.9371***	8.3848***	8.7125***	5.8127***
	(0.0658)	(0.0952)	(0.0927)	(0.2328)
Observations	854	854	854	780
R-squared	0.3092	0.3842	0.4386	0.5421

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 3 presents the results of four additional regression models where the dependent variable is still current health expenditure (% of GDP) and the independent variable is CO_2. Additional control variables include the real effective exchange rate index as a proxy of international trade, the age dependency ratio as a proxy of social and economic factors, out-of-pocket expenditure as a proxy of health system efficiency, and population as a proxy of country size. In model 1, climate risk is positively associated with current health expenditure after controlling for the effect of inflation, GDP growth, government expenditure on education, and the real effect of the exchange rate, with a coefficient of 0.0014. The result, statistically significant at the 1% level, indicates a higher positive correlation between climate risk and current health expenditure. In addition to model 1, model 2 controls for the age dependency ratio. The result is consistent with model 1; climate risk remains positively associated with current health expenditure, which is still statistically significant at 1%. Beyond model 2, model 3 and model 4 add out-of-pocket expenditure of health expenditure and population as additional control variables, respectively. The results are consistent with the previous models, emphasizing the positive relationship between climate risk and current health expenditure at a 1% level.

Table 3: An additional 4 regression models were added to Table 2, analyzing current health expenditure and showing a strong positive correlation with climate risk. Table 3 highlights the paper's findings by demonstrating how climate risk is the factor affecting current health expenditure while controlling for 7 other factors.

	(1)	(2)	(3)	(4)
VARIABLES	He	He	He	He
CO_2	0.0014***	0.0014***	0.0013***	0.0017***
	(0.0001)	(0.0001)	(0.0001)	(0.0002)
Inf	-0.2136***	-0.1938***	-0.1509***	-0.1479***
	(0.0239)	(0.0252)	(0.0230)	(0.0225)
GDP Growth	-0.0955***	-0.0898***	-0.0911***	-0.0976***
	(0.0188)	(0.0182)	(0.0156)	(0.0161)
Gov Exp Edu	0.6031***	0.5440***	0.4180***	0.3650***
	(0.0409)	(0.0412)	(0.0434)	(0.0477)
Exchange Rate	-0.0146***	-0.0107**	-0.0036	-0.0029
	(0.0044)	(0.0045)	(0.0040)	(0.0039)
Age Dep		0.0539***	0.0551***	0.0650***
		(0.0115)	(0.0094)	(0.0106)
Out of Pocket			-0.0624***	-0.0603***
			(0.0063)	(0.0062)
Population				-0.0000**
				(0.0000)
Constant	7.2010***	4.3190***	5.4516***	5.2180***
	(0.4482)	(0.8111)	(0.7137)	(0.7269)
Observations	702	702	699	699
R-squared	0.5356	0.5549	0.6163	0.6198

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Robustness was further established by using the ND-GAIN index as an alternative proxy of climate risk in OECD countries (Appendix 2 and Appendix 3). Results remained consistent with those reported in the main specification, indicating a positive relationship between health care expenditure and climate risk. Moreover, to cross-check the validity of our findings, additional measures were conducted: re-estimating the models excluding the U.S. and other outliers, applying log transformations to CO_2 emissions, and introducing lag structures to capture delayed effects. Across these specifications, the results remain consistent with those previously reported.

Lastly, it is also important to note that our primary objective was not to interpret the coefficients of each control variable individually, but rather to assess whether the effect of CO_2 emissions on health expenditure remains robust once different sets of controls are introduced.

Conclusion

This paper examines the impact of climate risk on health expenditure, highlighting the significant strain it places on healthcare systems and emphasizing the need for sustainable and resilient reforms. Among the 38 OECD countries, eight exhibit a significant direct relationship between rising climate risk, carbon emissions, and health expenditure. The overall relationship between climate risk and healthcare spending is positive across all OECD nations, which points out the extent of impact directed by these eight countries, underscoring the urgency of addressing climate-related health costs. The findings are further strengthened, evidenced by controlling for many variables. The paper urges policymakers to invest in healthcare infrastructure that can withstand extreme weather events, implement policies to reduce climate-induced illnesses, and integrate climate risk considerations into healthcare budgeting. Future research should explore the long-term economic implications of climate-related health expenditures, including their effects on government debt, insurance systems, and private healthcare spending. Additionally, further studies should assess the effectiveness of climate adaptation policies in mitigating healthcare costs and examine country-specific variations in climate health dynamics. A deeper understanding of these relationships will help develop more sustainable and adaptive healthcare financing strategies in response to increasing climate risk.

Acknowledgments

I want to express my gratitude to Prof. Abdullah Yalaman, Eskisehir Osmangazi University, Türkiye, and my parents for their continuous support.

Appendix

Appendix 1.

Variable Name	Definition	Source
He (Current health expenditure (% of GDP))	Level of current health expenditure expressed as a percentage of GDP. Estimates of current health expenditures include healthcare goods and services consumed during each year. This indicator does not include capital health expenditures such as buildings, machinery, IT, and stocks of vaccines for emergencies or outbreaks.	WDI: World Bank
CO_2 (Climate Risk)	Annual CO_2 emissions - Annual total emissions of carbon dioxide (CO_2), excluding land-use change, measured in million tonnes. CO_2 is measured in million tons per capita	Our World and Data
ND-GAIN (Climate Risk)	A measure of a country's vulnerability to climate change and its readiness to adapt. The index is scaled from 0 (highest risk, least prepared) to 100 (lowest risk, most prepared).	University of Notre Dame
Inf (Inflation, consumer price index (annual %))	Inflation as measured by the consumer price index reflects the annual percentage change in the cost to the average consumer of acquiring a basket of goods and services that may be fixed or changed at specified intervals, such as yearly. The Laspeyres formula is generally used.	WDI: World Bank
GDP Growth (GDP growth (annual %))	Annual percentage growth rate of GDP at market prices based on constant local currency. Aggregates are based on constant 2015 prices, expressed in U.S. dollars. GDP is the sum of gross value added by all resident producers in the economy, plus any product taxes, minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources.	WDI: World Bank
Gov Exp Edu (Government expenditure on education, total (% of GDP))	General government expenditure on education (current, capital, and transfers) is expressed as a percentage of GDP. It includes expenditure funded by transfers from international sources to the government. General government usually refers to local, regional, and central governments.	WDI: World Bank
Exchange Rate (Real effective exchange rate index (2010 = 100))	The real effective exchange rate is the nominal effective exchange rate (a measure of the value of a currency against a weighted average of several foreign currencies) divided by a price deflator or index of costs.	WDI: World Bank
Age Dep (Age dependency ratio (% of working-age population))	Age dependency ratio is the ratio of dependents–people younger than 15 or older than 64–to the working-age population–those ages 15-64. Data are shown as the proportion of dependents per 100 working-age population.	WDI: World Bank
Out of Pocket (Out-of-pocket expenditure (% of current health expenditure))	Share of out-of-pocket payments of total current health expenditures. Out-of-pocket payments are spending on health directly out-of-pocket by households.	WDI: World Bank
Population (persons)	Population by country, available from 10,000 BCE to 2100, based on data and estimates from different sources	Our World and Data

Appendix 2: Robustness check regression models analyzing current health expenditure reveal a strong positive association with climate risk (ND-GAIN index), providing evidence that changes in health expenditure are directly linked to climate risk.

	(1)	(2)	(3)	(4)
VARIABLES	He	He	He	He
ND-GAIN	0.1906***	0.1636***	0.2162***	0.2140***
	(0.0369)	(0.0344)	(0.0320)	(0.0408)
Inf		-0.2391***	-0.1003	-0.1018*
		(0.0868)	(0.0613)	(0.0601)
GDP Growth			-0.2883***	-0.2863***
			(0.0545)	(0.0495)
Gov Exp Edu				0.0241
				(0.2306)
Constant	-2.4803	-0.4613	-5.1238**	-5.1021**
	(2.3026)	(2.1463)	(2.0423)	(2.0899)
Observations	38	38	38	38
R-squared	0.2876	0.3242	0.4647	0.4648

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Appendix 3: Robustness check with 4 additional regression models on current health expenditure shows a strong positive correlation with climate risk (ND-GAIN index), providing evidence that variations in health expenditure are directly associated with climate risk.

	(1)	(2)	(3)	(4)
VARIABLES	He	He	He	He
ND-GAIN	0.2070***	0.2046***	0.1476***	0.1679***
	(0.0399)	(0.0483)	(0.0528)	(0.0337)
Inf	-0.1321	-0.1131	-0.2243	-0.3990*
	(0.2103)	(0.2541)	(0.2251)	(0.2016)
GDP Growth	-0.2630***	-0.2604***	-0.2952***	-0.2248***
	(0.0522)	(0.0589)	(0.0556)	(0.0508)
Gov Exp Edu	0.0417	0.0399	0.0596	0.3879
	(0.2641)	(0.2632)	(0.2777)	(0.2343)
Exchange Rate	-0.0034	-0.0020	0.0041	-0.0205
	(0.0395)	(0.0451)	(0.0446)	(0.0161)
Age Dep		0.0098	-0.0027	-0.0438
		(0.0636)	(0.0665)	(0.0480)
Out of Pocket			-0.0969	-0.0494
			(0.0674)	(0.0460)
Population				0.0000***
				(0.0000)
Constant	-4.2275	-4.7392	0.6139	0.7505
	(4.0259)	(6.1673)	(5.1563)	(3.6034)
Observations	34	34	34	34
R-squared	0.3832	0.3837	0.4405	0.7959

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

References

1. Cantarero, D.; Lago-Peñas, S. The Determinants of Health Care Expenditure: A Reexamination. Appl. Econ. Lett. 2010, 17 (7), 723–726. DOI: 10.1080/13504850802297873.
2. Martín, J. J. M.; Puerto Lopez del Amo Gonzalez, M.; Cano Garcia, M. D. Review of the Literature on the Determinants of Healthcare Expenditure. Appl. Econ. 2011, 43 (1), 19–46. DOI: 10.1080/00036841003689754.
3. Cheng, C.; Ren, X.; Zhang, M.; Wang, Z. The Nexus among CO_2 Emission, Health Expenditure and Economic Development in the OECD Countries: New Insights from a Cross-Sectional ARDL Model. Environ. Sci. Pollut. Res. 2024, 31 (11), 16746–16769. DOI: 10.1007/s11356-024-15827-2.
4. Dritsaki, M.; Dritsaki, C. The Relationship between Health Expenditure, CO_2 Emissions, and Economic Growth in G7: Evidence from Heterogeneous Panel Data. J. Knowl. Econ. 2024, 15 (1), 4886–4911. DOI: 10.1007/s13132-019-00634-0.
5. Gbesemete, K. P.; Gerdtham, U. G. Determinants of Health Care Expenditure in Africa: A Cross-Sectional Study. World Dev. 1992, 20 (2), 303–308. DOI: 10.1016/0305-750X(92)90108-8.
6. Horvey, S. S. Towards the Cost of Health in Africa: Examining the Synergistic Effect of Climate Change and Renewable Energy on Health Expenditure. Air Qual., Atmos. Health 2024, 1–23. DOI: 10.1007/s11869-024-01651-x.
7. Alimi, O. Y.; Ajide, K. B.; Isola, W. A. Environmental Quality and Health Expenditure in ECOWAS. Environ. Dev. Sustain. 2020, 22, 5105–5127. DOI: 10.1007/s10668-019-00460-4.
8. Furuoka, F.; Fui Yee, B. L.; Kok, E.; Hoque, M. Z.; Munir, Q. What Are the Determinants of Health Care Expenditure? Empirical Results from Asian Countries. Sunway Acad. J. 2011, 8, 12–25.
9. Gatzert, N.; Reichel, P. Awareness of Climate Risks and Opportunities: Empirical Evidence on Determinants and Value from the U.S. and European Insurance Industry. Geneva Pap. Risk Insur. Issues Pract. 2022, 47, 5–26. DOI: 10.1057/s41288-021-00227-5.
10. Leppänen, S.; Solanko, L.; Kosonen, R. The Impact of Climate Change on Regional Government Expenditures: Evidence from Russia. Environ. Resour. Econ. 2017, 67, 67–92. DOI: 10.1007/s10640-015-9977-y.
11. Gao, P.; Rong, Y.; Cao, Y.; Zhang, Q.; Sun, H. Regional Climate Risks and Government Education Expenditure: Evidence from China. Front. Energy Res. 2024, 12, 1374065. DOI: 10.3389/fenrg.2024.1374065.
12. Chen, L.; Zhuo, Y.; Xu, Z.; Xu, X.; Gao, X. Is Carbon Dioxide (CO_2) Emission an Important Factor Affecting Healthcare Expenditure? Evidence from China, 2005–2016. Int. J. Environ. Res. Public Health 2019, 16 (20), 3995. DOI: 10.3390/ijerph16203995.
13. Ullah, I.; Ali, S.; Shah, M. H.; Yasim, F.; Rehman, A.; Al-Ghazali, B. M. Linkages between Trade, CO_2 Emissions and Healthcare Spending in China. Int. J. Environ. Res. Public Health 2019, 16(21), 4298. DOI: 10.3390/ijerph16214298.
14. Islam, M. S.; Islam, A. H. M. S.; Sato, M. Nexus between Climatic Extremes and Household Expenditures in Rural Bangladesh: A Nationally Representative Panel Data Analysis. Asia-Pac. J. Reg. Sci. 2023, 7 (2), 355–379. DOI: 10.1007/s41685-022-00266-3.
15. Bae, S. M.; Masud, M. A. K.; Rashid, M. H. U.; Kim, J. D. Determinants of Climate Financing and the Moderating Effect of Politics: Evidence from Bangladesh. Sustain. Account. Manag. Policy J. 2022, 13 (1), 247–272. DOI: 10.1108/SAMPJ-05-2021-0197.
16. Islam, M. F.; Debnath, S.; Das, H.; Hasan, F.; Sultana, S.; Datta, R.; Halimuzzaman, M.; et al. Impact of Rapid Economic Development with Rising Carbon Emissions on Public Health and Healthcare Costs in Bangladesh. J. Angiotherapy 2024, 8 (7), 1–9.
17. Apergis, N.; Gupta, R.; Lau, C. K. M.; Mukherjee, Z. U.S. State-Level Carbon Dioxide Emissions: Does It Affect Health Care Expenditure? Renew. Sustain. Energy Rev. 2018, 91, 521–530. DOI: 10.1016/j.rser.2018.03.103.
18. Gündüz, M. Healthcare Expenditure and Carbon Footprint in the USA: Evidence from Hidden Cointegration Approach. Eur. J. Health Econ. 2020, 21 (5), 801–811. DOI: 10.1007/s10198-020-01174-z.
19. Wang, K. M. Health Care Expenditure and Economic Growth: Quantile Panel-Type Analysis. Econ. Model. 2011, 28(4), 1536–1549. DOI: 10.1016/j.econmod.2011.02.008.
20. Hartwig, J.; Sturm, J. E. Robust Determinants of Health Care Expenditure Growth. Appl. Econ. 2014, 46 (36), 4455–4474. DOI: 10.1080/00036846.2014.964829.
21. Ampon-Wireko, S.; Zhou, L.; Xu, X.; Dauda, L.; Mensah, I. A.; Larnyoh, E.; Baah Nketia, E. The Relationship between Healthcare Expenditure, CO_2 Emissions and Natural Resources: Evidence from Developing Countries. J. Environ. Econ. Policy 2022, 11 (3), 272–286. DOI: 10.1080/21606544.2022.2035224.
22. O'Neill, B. C.; Jiang, L.; Kc, S.; Fuchs, R.; Pachauri, S.; Laidlaw, E. K.; Ren, X. The Effect of Education on Determinants of Climate Change Risks. Nat. Sustain. 2020, 3 (7), 520–528. DOI: 10.1038/s41893-020-0512-y.
23. Chaabouni, S.; Saidi, K. The Dynamic Links between Carbon Dioxide (CO_2) Emissions, Health Spending and GDP Growth: A Case Study for 51 Countries. Environ. Res. 2017, 158, 137–144.
24. World Bank. World Development Indicators. https://databank.worldbank.org/source/world-development-indicators.
25. Our World in Data. https://ourworldindata.org.

Author

Albert Luo is currently a sophomore at The Webb Schools, located in Claremont, California.

Study of The Cycloidal Curves and The Application in Hydraulic Motor Design

Jialin Dong

TVT Community Day School, 5 Federation Way, Irvine, CA 92603, USA; kelvindong66@gmail.com

ABSTRACT: The research described in this paper is a part of the design and development of a compact and high-torque hydraulic motor for robotic arms. Traditional motors are too bulky to be installed on robotic arms. This paper presents new designs of hydraulic motors based on cycloidal curves. It presents the mathematically detailed generation of both epicycloidal and hypocycloidal curves, including the standard, shortened, and modified cycloids, for cycloidal gears and corresponding pin gears. The innovative design of hydraulic epicycloidal and hypocycloidal motors (also known as orbital motors) was described, including designs of gears and the oil distribution system. The comparison with traditional orbital motors is discussed, and the advantages of the new design, including compact size and high precision, are highlighted. A hydraulic motor for a subsea robotic arm was designed as a real industrial design case. Compact size, high output torque, and smooth spinning at low speeds were needed. The method presented was used to make an orbital motor with seven teeth on the inner gear. The designed motor was installed on a subsea robotic arm and has been operating in a subsea environment for over a year. This design case completely proves that the theory and design method proposed here are effective.

KEYWORDS: Engineering Mechanics, Mechanical Engineering, Robotic Arm, Cycloidal Gear, Hydraulic Motor, Robotics.

Introduction

The cycloids, the curves traced by a point on a rolling circle, have captivated mathematicians since the Renaissance. Although their properties were hinted at in antiquity, serious study began in the 17th century. Galileo Galilei (1599) is often credited with naming the cycloids and attempting to calculate the area under one arch, though his results were approximate.[1] The curve's true mathematical exploration flourished during the "century of genius": Blaise Pascal (1658) solved key problems related to its area and centroid, while Christiaan Huygens (1659) discovered its property, using it to design pendulum clocks with improved accuracy. The cycloids became a battleground for calculus pioneers—Johann Bernoulli, Jakob Bernoulli, Gottfried Leibniz, and Isaac Newton—who tackled the Brachistochrone problem (1697), proving the cycloids' optimality as the "curve of fastest descent." The cycloids' applications in physics, engineering, and mathematics cemented their legacy as a cornerstone of classical mechanics and calculus. The rich history reflects both the beauty of pure mathematics and its profound utility.[2]

Hydraulic manipulators are popularly used in subsea applications. Due to strict limitations on size and weight, joint actuators are required to be compact and powerful. Many companies in this area encountered the same issue: hydraulic low-speed-high-torque (LSHT) motor for the wrist joints of their robotic arms were overly bulky and asymmetrical, causing operational inconvenience, visual obstruction, and limiting their usage scenarios. The motivation of this research is to use cycloids to find a better design of hydraulic motors for wrist joints of robotic arms.

In cycloids' application, the cycloidal pinion gear transmission system (Cycloidal Drive Systems)[3] utilizes the geometric properties of epicycloid and hypocycloid, achieving high-precision power transmission through precise meshing mechanisms. In the field of hydraulic motors and reducers, the cycloidal pinion gear transmission system is widely used due to its high efficiency and flexibility. For example, in single-stage reducers,[4,5] their output efficiency can reach over 98%, significantly reducing energy consumption and enhancing production efficiency. This paper explores the theory of cycloids and cycloidal transmission, which would be used in the innovation of hydraulic motor design.

The core of cycloidal pinion transmission is the great role of mathematical modeling of cycloidal tooth shape in gear meshing (mesh). Characteristics of cycloidal transmission are as follows.[6,7]

1) Compact and Lightweight Design

The rotor (pin gear) and stator (cycloidal gear) utilize cycloidal profiles. The rotor performs planetary motion via an eccentric shaft, eliminating the need for multi-stage transmissions, drastically reducing size. Its compact layout allows lighter weight compared to gear or piston motors of equivalent power, making them ideal for space-constrained systems (e.g., AGVs, robotic joints [8]).

2) Structural symmetry

To meet the bidirectional rotation requirements of hydraulic motors, the cycloidal gear and pin gear are adopted with a symmetrical structure. Reversing fluid flow direction enables easy forward/reverse switching without additional mechanisms. Through the precise cooperation between the eccentric cycloidal gear and the pin gear, the stable output of the drive is ensured. Speed is adjustable from near-zero to hundreds of rpm and adaptable to diverse operational needs.

DOI: 10.36838/v8i2.27

3) Low speed and high torque characteristics

The geometric properties of cycloidal gears enable significant torque generation even at low speeds (high torque-to-volume ratio), ideal for applications requiring heavy-load, low-speed operation (e.g., cranes, excavator slewing mechanisms). Continuous meshing of cycloidal teeth minimizes output pulsations, ensuring stable operation even at extremely low speeds (e.g., 1-2 rpm) without "crawling" effects.

4) High Mechanical Efficiency and Durability

Multiple contact points (typically 6-8) during meshing ensure even pressure distribution, reducing localized wear. Rolling friction dominance further enhances energy efficiency. Critical components (e.g., rotor, stator) use hardened steel or composites for wear resistance. Hydraulic oil provides direct lubrication, minimizing the frequency of maintenance needed.

Although the geometric problem of the cycloidal gear is based on the parametric equation of a circle (positive and negative cosine trigonometric functions), its correct mathematical derivation becomes a great difficulty and challenge due to its dynamic coordinate transformation. In the second part of this paper, the mathematical derivation of cycloids is provided with details in standard, shortened, and modified versions, which leads to a whole theory to generate cycloid curves for gear design. The third part describes the way to design a cycloidal motor (also known as an orbital motor), followed by a design case study of an orbital motor based on a hypocycloid. Then, the test results of the designed motor are presented and discussed.

■ Mathematical Modeling of Cycloids

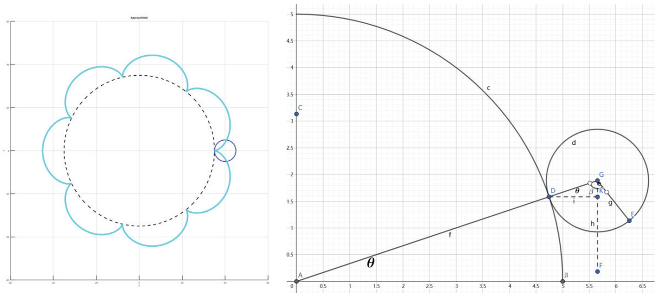
Basic Cycloids:

Firstly, the mathematical cycloidal model needs to be established. The Epicycloidal and Hypocycloidal curves are mathematically described in the following two parts.

1) Epicycloids

As shown in Figure 1, let the big circle (shown in the quarter) be the base one with radius R, and the small circle is the rolling one with radius r. The parametric equations (i.e., the coordinates of a point) for the base circle can be written as:

$$\begin{cases} x = R \cdot \cos\theta \\ y = R \cdot \sin\theta \end{cases} \tag{1}$$



(a) (b)

Figure 1: The geometric drawing of an epicycloidal curve. (a) The definition of a standard epicycloidal curve. (b) An enlarged drawing to describe the mathematical derivation.

Where θ is the angle between the line connecting the point to the origin and the positive x-axis, the coordinates of the rolling circle's center are:

$$\begin{cases} x_r = (R + r) \cdot \cos\theta \\ y_r = (R + r) \cdot \sin\theta \end{cases} \tag{2}$$

When the rolling circle rolls on the base circle by an angle θ, it rotates around its center by an angle. Because the rolling has no slipping, the arc lengths traced on the two circles are equal, i.e.,

$$R = z \cdot r \tag{3}$$

Here, z is an integer, which determines the number of petals of the cycloid, which is the number of teeth in a pin gear. Therefore, the coordinates of the reference point on the rolling circle are:

$$\begin{cases} x_e = x_r + r \cdot \sin\alpha \\ y_e = y_r - r \cdot \cos\alpha \end{cases} \tag{4}$$

Where,

$$\alpha = \angle EGF = \angle AGE - \angle AGF = \beta - (\tfrac{\pi}{2} - \theta) \tag{5}$$

Replacing x_r, y_r, and r by Eqns (2) and (3),

$$\begin{cases} x_e = r(z + 1) \cdot \cos\theta + r \cdot \sin(\beta - (\tfrac{\pi}{2} - \theta)) \\ y_e = r(z + 1) \cdot \sin\theta - r \cdot \cos(\beta - (\tfrac{\pi}{2} - \theta)) \end{cases} \tag{6}$$

Also, the arc BD is the same as arc DE, so

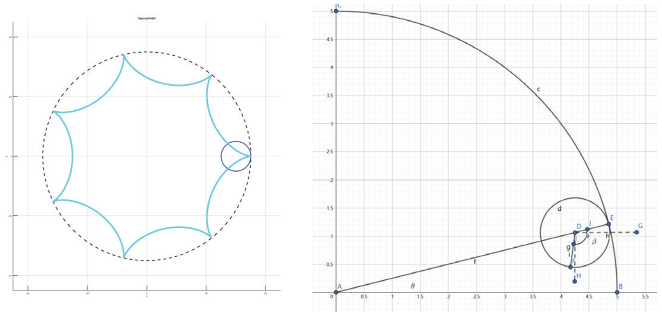
$$\theta \cdot R = \beta \cdot r \tag{7}$$

Finally, the expression of a standard epicycloid can be derived

$$\begin{cases} x_e = r(z + 1) \cdot \cos\theta - r \cdot \cos[(1 + z)\theta] \\ y_e = r(z + 1) \cdot \sin\theta - r \cdot \sin[(1 + z)\theta] \end{cases} \tag{8}$$

2) Hypocycloids

As shown below, let the dashed line represent the base circle with radius R, and the blue circle represent the rolling circle with radius r. The parametric equations (i.e., coordinates of a point) for the base circle can be expressed as Eqn (1):



(a) (b)

Figure 2: The geometric drawing of a hypocycloidal curve. (a) The definition of a standard hypocycloidal curve. (b) An enlarged drawing to describe the mathematical derivation.

The coordinates of the rolling circle's center are:

$$\begin{cases} x_r = (R - r) \cdot \cos\theta \\ y_r = (R - r) \cdot \sin\theta \end{cases} \tag{9}$$

When the rolling circle has rolled along the base circle by an angle θ, it rotates around its center by an angle. Since the rolling has no slipping, the arc lengths traced on both circles are equal, i.e., R and r have the same relation shown in Eqn(3),

Here, z is an integer that determines the number of lobes (petals) of the hypocycloid, corresponding to the number of teeth in the pin gear.

Thus, the coordinates of a reference point on the rolling circle are:

$$\begin{cases} x_h = x_r - r \cdot \sin\gamma \\ y_h = y_r - r \cdot \cos\gamma \end{cases} \quad (10)$$

Where,

$$\gamma = \angle HDF = \angle FDE - \angle HDE = \beta - (\tfrac{\pi}{2} + \theta) \quad (11)$$

Since the rolling circle moves on the base circle without sliding, the lengths of the arc FE and arc BE are the same, Eqn (12) is derived

$$R \cdot \theta = r \cdot \beta \quad (12)$$

After substituting Eqs (9, 11, 12) into Eqn (10), Eqn (13) can be obtained.

$$\begin{cases} x_h = r(z-1) \cdot \cos\theta - r \cdot \sin[\beta - (\tfrac{\pi}{2} + \theta)] \\ y_h = r(z-1) \cdot \sin\theta - r \cdot \cos[\beta - (\tfrac{\pi}{2} + \theta)] \end{cases} \quad (13)$$

This simplifies a standard hypocycloid to be expressed as:

$$\begin{cases} x_h = r(z-1) \cdot \cos\theta + r \cdot \cos[(z-1) \cdot \theta] \\ y_h = r(z-1) \cdot \sin\theta - r \cdot \sin[(z-1) \cdot \theta] \end{cases} \quad (14)$$

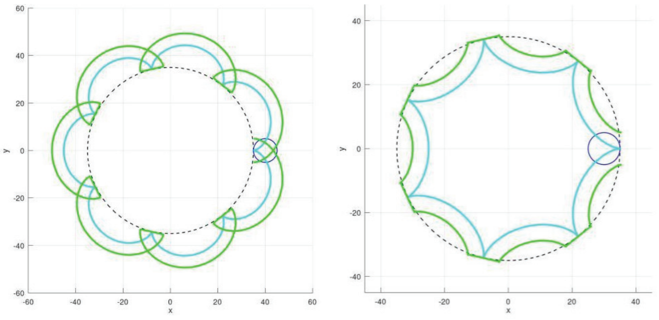
Revision and Modification:

The standard cycloid has a serious flaw--it contains mathematically discontinuous points (non-differ-entiable points). To create a pin gear suitable for transmission, it's necessary to apply a displacement (offset) to create space for the rollers on the pin teeth to operate.

The method of displacement involves extending or shrinking the curve along its normal direction by a specific distance. Specifically, by calculating the differentials {dx/dθ, dy/dθ} at a given point, the proportional components in the x and y directions can be determined. The displacement is then applied according to the ratio of these components over the total displacement distance, as follows:

$$\begin{cases} l_x = l \cdot \dfrac{\frac{dx}{d\theta}}{\sqrt{\frac{dx}{d\theta}^2 + \frac{dy}{d\theta}^2}} \\ l_y = l \cdot \dfrac{\frac{dy}{d\theta}}{\sqrt{\frac{dx}{d\theta}^2 + \frac{dy}{d\theta}^2}} \end{cases} \quad (15)$$

However, because the standard cycloid has non-differentiable (discontinuous) points, it is impossible to calculate the displacement at these locations (see Figure 3).



(a) (b)

Figure 3: Modification of standard cycloidal curves: (a) Standard epicycloidal curves, and (b) Standard hypocycloidal curves.

Thus, the standard cycloidal curves need to be revised. By replacing the in the subtracted term with $\eta \cdot r$, where η is a percentage less than 100% (i.e., using a point inside the rolling circle instead of on its circumference), a shortened epicycloid can be generated. In conclusion, the equations of the shortened epicycloids and hypocycloids can be derived as Eqs (16) and (17)

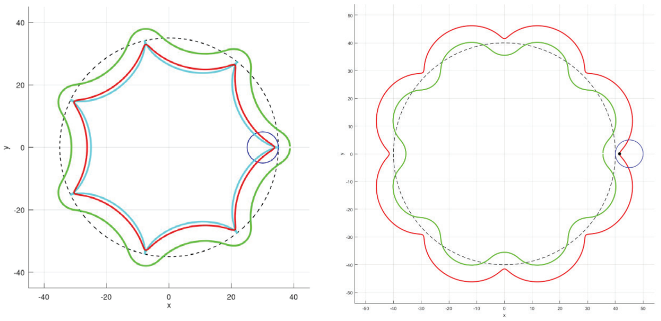
$$\begin{cases} x_{es} = r(z+1) \cdot \cos\theta - \eta \cdot r \cdot \cos[(1+z)\theta] \\ y_{es} = r(z+1) \cdot \sin\theta - \eta \cdot r \cdot \sin[(1+z)\theta] \end{cases} \quad (16)$$

$$\begin{cases} x_{hs} = r(z-1) \cdot \cos\theta + \eta \cdot r \cdot \cos[(z-1) \cdot \theta] \\ y_{hs} = r(z-1) \cdot \sin\theta - \eta \cdot r \cdot \sin[(z-1) \cdot \theta] \end{cases} \quad (17)$$

Next, the shortened cycloid in red in the graph is scaled equidistantly to form the green modified cycloids. This process uses proportional scaling: for a segment on the cycloid, its x- and y-direction components are scaled according to their proportional ratios given by Eqns (18) and (19).

$$\begin{cases} x_{em} = x_{es} - l \cdot \dfrac{\frac{dx_{es}}{d\theta}}{\sqrt{\frac{dx_{es}^2}{d\theta} + \frac{dy_{es}^2}{d\theta}}} = r(z+1) \cdot \cos\theta - \eta \cdot r \cdot \cos[(1+z)\theta] - l \cdot \dfrac{\frac{dx_{es}}{d\theta}}{\sqrt{\frac{dx_{es}^2}{d\theta} + \frac{dy_{es}^2}{d\theta}}} \\ y_{em} = y_{es} - l \cdot \dfrac{\frac{dy_{es}}{d\theta}}{\sqrt{\frac{dx_{es}^2}{d\theta} + \frac{dy_{es}^2}{d\theta}}} = r(z+1) \cdot \sin\theta - \eta \cdot r \cdot \sin[(1+z)\theta] - l \cdot \dfrac{\frac{dy_{es}}{d\theta}}{\sqrt{\frac{dx_{es}^2}{d\theta} + \frac{dy_{es}^2}{d\theta}}} \end{cases} \quad (18)$$

$$\begin{cases} x_{hm} = x_{hs} + l \cdot \dfrac{\frac{dx_{hs}}{d\theta}}{\sqrt{\frac{dx_{hs}^2}{d\theta} + \frac{dy_{hs}^2}{d\theta}}} = r(z-1) \cdot \cos\theta + \eta \cdot r \cdot \cos[(z-1) \cdot \theta] + l \cdot \dfrac{\frac{dx_{hs}}{d\theta}}{\sqrt{\frac{dx_{hs}^2}{d\theta} + \frac{dy_{hs}^2}{d\theta}}} \\ y_{hm} = y_{hs} + l \cdot \dfrac{\frac{dy_{hs}}{d\theta}}{\sqrt{\frac{dx_{hs}^2}{d\theta} + \frac{dy_{hs}^2}{d\theta}}} = r(z-1) \cdot \sin\theta - \eta \cdot r \cdot \sin[(z-1) \cdot \theta] + l \cdot \dfrac{\frac{dy_{hs}}{d\theta}}{\sqrt{\frac{dx_{hs}^2}{d\theta} + \frac{dy_{hs}^2}{d\theta}}} \end{cases} \quad (19)$$



(a) (b)

Figure 4: Modification of shortened cycloidal curves: (a) Shortened epicycloidal curves, and (b) Shortened hypocycloidal curves.

Hydraulic Orbital Motor Design

The author used the mathematical program in Octave to generate the cycloidal gear and pin gear's shape, and CAD software to make a 3D model of the motor. Before giving the details on how to design two types of motors in the following

DOI: 10.36838/v8i2.27

two parts, three concepts need to be defined: rotor, stator, and float stator.

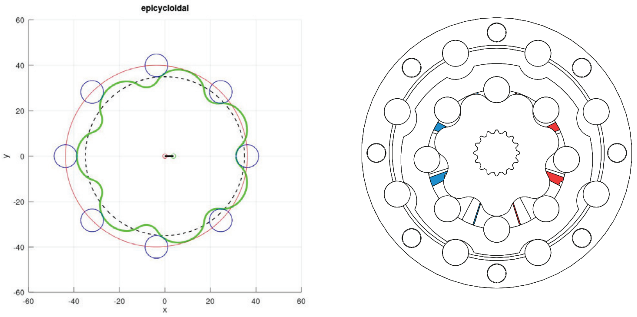
The rotor is the inner gear, which can spin along the central axis.

A stator is a part that cannot spin and is normally fixed to the housing.

The float stator is the outer gear that is not spinning but movable to adjust the contact position with the rotor so that mechanical transmission can be achieved.

Epicycloidal Gear and Pin Gear Design:

The epicycloidal gear is an inner gear, i.e., rotor. The design is based on Eqn (18). The equation can be used to generate an epicycloidal curve with four parameters: radius of rolling circle *r*, integer ratio of base circle to rolling circle *z* (i.e., the number of petals), shorten coefficient η , and displacement distance *l*. The curve, which is shown as the green curve in Figure 5(a), forms the contact surface of a cycloidal gear. For an epicycloidal hydraulic motor, the cycloidal gear is the inner gear, while the pin gear is the outer gear, which can be seen in Figure 5(b).



(a) (b)

Figure 5: Epicycloid-based orbital motor design. (a) Gear curves generation. (b) Use the generated curve to design a motor in 3D CAD software.

The pin gear is the outer gear, i.e., the float stator, and is formed by several pins, which are tangent to the epicycloidal curve. Also, it is eccentric to the curve, and the eccentric offset is the shortened radius to form the curve *0102* = $\eta \cdot r$. Therefore, there are z+1 circular teeth shown as the blue circles in Figure 5(a) located on a circle eccentric to the origin, which is shown as the red circle. These cylinders are just tangent to the epicycloid. There are two methods to design the pin gear after determining the epicycloidal curve: mathematical and engineering methods.

Mathematical Method. The small red and green circles are the center of the red circle and the green curve, respectively. The red circle is

$$\begin{cases} x_{epb} = (R + r) * \cos\theta - \eta \cdot r \\ y_{epb} = (R + r) * \sin\theta \end{cases} \qquad (20)$$

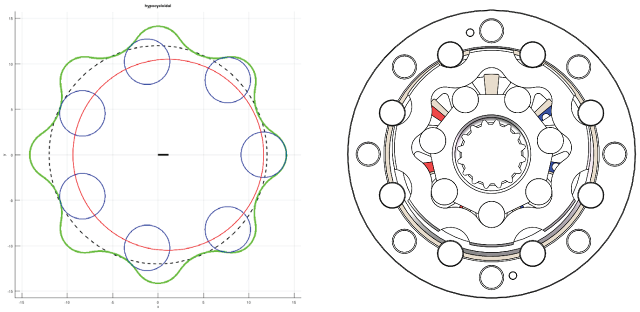
The *i*-th pin on the pin gear can be expressed as

$$\begin{cases} x_{epc\_i} = l \cdot \cos(\theta) + (R + r) \cdot \cos(2\pi/(z + 1) \cdot (i - 1)) - \eta \cdot r \\ y_{epc\_i} = l \cdot \sin\theta + (R + r) \cdot \sin(2 \cdot \pi/(z + 1) \cdot (i - 1)) \end{cases} \qquad (21)$$

Engineering Method. In CAD software, place cylinders equispaced with a radius of *l* tangent to the epicycloid from outside, which is shown in Figure 5(b).

Hypocycloidal Gear and Pin Gear Design:

Similar to the epicycloidal gear, the hypocycloidal gear design is based on Eqn (19). However, it is the outer gear, i.e., the float stator. The equation can be used to generate a hypocycloid with four parameters: radius of rolling circle *r*, integer ratio of base circle to rolling circle *z* (i.e., the number of petals), shortening coefficient η, and displacement distance *l*. The curve, which is shown as the green curve in Figure 6(a), forms the contact surface of a cycloidal gear.



(a) (b)

Figure 6: Hypocycloid-based orbital motor design. (a) Gear curves generation. (b) Use the generated curve to design a motor in 3D CAD software.

The pin gear is formed by several pins, which are tangent to the hypocycloid from inside. It is eccentric to the curve, and the eccentric offset is the shortened radius to form the curve *0102* = $\eta \cdot r$. Differently, the pin gear in a hypocycloid-based orbital motor is a rotor. There are *z*+1 circular teeth shown as the blue circles in Figure 6(a) located on a circle eccentric to the origin, which is shown as the red circle. These cylinders are just tangent to the epicycloid. There are two methods to design a pin gear after determining the hypocycloid: mathematical and engineering methods.

Mathematical Method. The small red and green circles are the center of the red circle and the green curve, respectively. The red circle is

$$\begin{cases} x_{hpb} = (R - r) * \cos\theta + \eta \cdot r \\ y_{hpb} = (R - r) * \sin\theta \end{cases} \qquad (22)$$

and the *i*-th pin on the pin gear can be expressed as

$$\begin{cases} x_{hpc\_i} = l \cdot \cos(\theta) + (R - r) \cdot \cos(2\pi/z \cdot (i - 1)) + \eta \cdot r \\ y_{hpc\_i} = l \cdot \sin\theta + (R - r) \cdot \sin(2 \cdot \pi/z \cdot (i - 1)) \end{cases} \qquad (23)$$

Engineering Method.

Similarly, in CAD software, place cylinders equispacedly with a radius tangent to the hypocycloid from the inner side, as shown in Figure 6(b).

Oil Distribution Design.

It can be seen from Figures 5 and 6 that a pair of rotor and float stator of both epicycloid and hypocy-cloid-based orbital motors creates cavities, each of which is formed by two pins, a segment of curve on the cycloid, and a segment of curve on the pin gear. The only difference is that the cycloidal curve segment of the hypocycloid base orbital motor is outside, while that of the epicycloid base orbital motor is inside.

When the rotors of both motors spin, half (if the number of cavities is even) or nearly half (if the number of cavities is odd) cavities tend to be enlarged, and the others are the contrary. For example, when the rotor of the epicycloid-based orbital motor shown in Figure 5 spins clockwise, the cavities with blue marks get smaller. The volume of a cavity turns to decreases when it crosses the upper half of the vertical bisector. Conversely, the cavity's volume increases when it crosses the lower half of the vertical bisector. Therefore, the oil distribution is easily implemented by the following theory.

Oil Distribution Theory: After deciding the direction to spin, the oil distribution system always fills the cavities, which tend to be larger with high-pressure oil, and allows the ones that tend to be smaller to output oil back to the low-pressure tank.

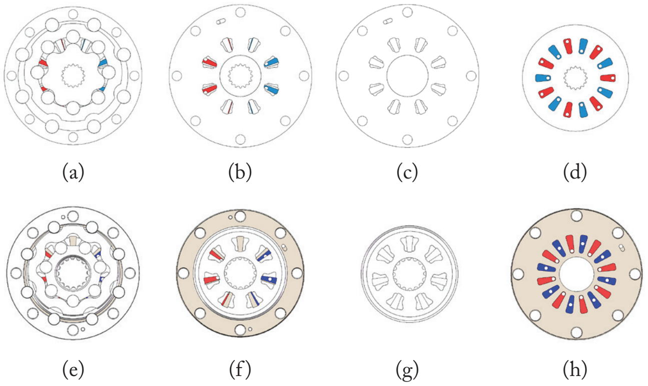


Figure 7: The oil distribution systems of epicycloid and hypocycloid base orbital motors. The first row is 3D models of an epicycloid-based orbital motor, while the second row is those of a hypocy-cloid-based one. (a) and (e) are epicycloid and hypocycloid-based orbital motors. (b) and (f) are the oil distribution systems of both motors. (c) and (g) are the oiling plates. (d) and (h) are the distributing plates.

The oil distribution system consists of two parts: the oiling plate and the distributing plate. The oiling plate is the one that contacts both gears and fills oil into the cavities. The distributing plate is the one that contacts and rotates relatively to the oiling plate. Since there is a relative rotation between these two plates, the cavities would either be filled with oil or output oil according to the relative angle.

Figure 7 shows the oil distribution systems of two types of motors. It can be found that the number of oiling holes is equal to the number of teeth of the pin gears. The number of distributing holes is twice the number of cycloidal gears. Table 1 shows a summary of the design issues of both motors.

Table 1: Summary of the structure of epicycloid and hypocycloid-based orbital motors.

Item	Epicycloid Base Motor	Hypocycloid-Based Motor
Inner Gear	Cycloidal Gear	Pin Gear
Outer Gear	Pin Gear	Cycloidal Gear
Oiling Plate	Stationery to Housing	Stationary to Rotor
Distributing Plate	Stationary to Rotor	Stationery to Housing
Number of Teeth, Inner Gear	z	z
Number of Teeth, Outer Gear	$z+1$	$z+1$
No. of Oiling Holes	$z+1$	z
No. of Distributing Holes	$2z$	$2(z+1)$

Comparing the oil distribution systems of both motors, it can be seen that the epicycloid-based orbital motor is more difficult to design. An oiling hole shall always be between two pins. However, the pin gear of the epicycloid base orbital motor is a floating part, which slides in the housing. Therefore, oiling holes should be sized and located to guarantee that it is always between two pins, even when the pin gear is sliding, which is a difficulty that does not exist in a hypocycloid-based orbital motor.

Design Case

After analyzing the mathematical model of two cycloidal curves and discussing the design of oil distribution systems, a motor can be easily designed and manufactured. Firstly, for a subsea manipulator, a hypocycloid-based orbital motor was determined to be designed and used. The limited size of such manipulators requires the diameter of our motor to be less than 110mm. After a brief sketch, a diameter of around 50mm for our base circle is determined. The design starts with the sketch and then iterates according to the design result. The specifications are listed in Table 2.

Assume the base circle radius is R = 25mm. The number of inner gear teeth is z = 7, and therefore one of the outer gear teeth is z+1 = 8. The radius of the rolling circle is r = R/z = 3.57mm. The author selects the eccentric distance r_s = 2.5mm, and thus the shortening coefficient is $\eta = r_s/r = 0.7$. To use standard bearing rollers, l = 5mm is determined. Figure 1 shows the design and machined parts.

Table 2: Specification parameters of the designed hypocycloid-based orbital motor.

Parameter	Symbol	Value
Radius of base circle	R	25mm
Radius of rolling circle	r	3.57mm
Shorten coefficient	η	0.7
Displace	l	5mm
Radius of roller	r_r	5mm
No. of teeth, pin gear	z	7
No. of teeth, cycloidal gear	$z+1$	8
Eccentric distance	0102	2.5mm
Thickness of gear	b	20mm
Volume displacement	V_Q	116ml/rev

The motor is assembled and tested on a dynamometer. The motor is installed on the dynamometer, and the spline shaft is connected to a magnetic adjustable load. The motor is powered by a pressure-controlled hydraulic power unit and controlled by a servo valve. The input pressure, spinning speed, and flow are all monitored. The test results are achieved and shown in Table 3. According to the design specification, the theoretical output torque is

$$\tau = \frac{P * V_Q}{2\pi} = 387.7\text{Nm}$$

DOI: 10.36838/v8i2.27

The test results show that the efficiency of the motor is

$$\eta_e = \frac{330}{387.7} = 85\%$$

The performance of the low-speed high-torque (LSHT) motor is good. The special structure makes each cavity change the status of filling and discharging by z*(z+1) times, which is equivalent to a z*(z+1):1 reducer. It can run smoothly when the speed is 3 rpm, which is very useful for extremely low-speed applications such as robotic joints.

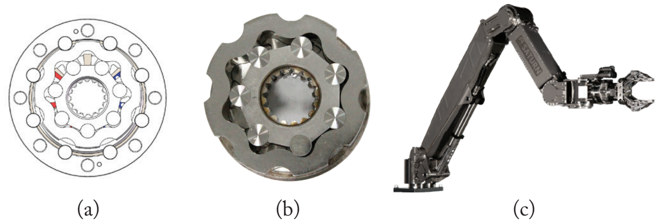

(a) (b) (c)

Figure 8: The design process. (a) the design in CAD software, (b) the machined and assembled gears, and (c) the subsea manipulator, the wrist of which is equipped with the hypocycloid-based orbital motor.

Table 3: Test results of the designed hypocycloid-based orbital motor.

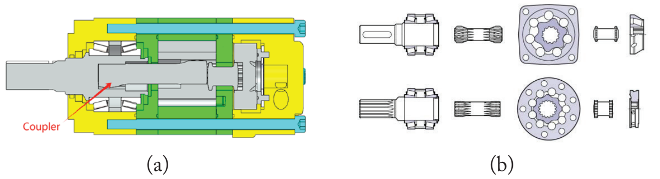
Displacement	Maximum Speed	Continuous Flow Rate/ Instantaneous Flow Rate	Continuous Pressure Instantaneous Pressure	Continuous Torque Instantaneous Torque	Maximum Power
116ml/rev	716 rpm	70L/min 80L/min	14MPa 21MPa	278Nm 330Nm	26.5HP

Discussion

Traditional orbital motors normally use an inner epicycloidal gear, a pin gear, and a spline coupler. The innovative application of a floating stator replacing a spline coupler makes the presented motors superior to traditional orbital motors. The main advantages are as follows:

1. The motor length in the axial direction is dramatically decreased. In a traditional motor, the outer gear and housing are the same part, which is fixed. As presented in the mathematical part, the inner gear needs to spin eccentrically if the outer gear is fixed. The traditional one needs to use a coupler with splines at each end, shown in Figure 9, which requires extra length (even more than double the length) in the axial direction.

2. Since one end of the coupler needs to spin with sliding movement in a traditional motor, the spline cannot be too tight, which means that the backlash is big, and the accuracy is not high enough for precise servo control.



(a) (b)

Figure 9: Traditional epicycloid-based orbital motor manufactured by Danfoss Power Solutions (US) Company.[9] (a) the crossing-section view, (b) the explosive view of coupler transmission with oil distribution plate.

On the other hand, the only drawback of the presented epicycloid base orbital motor is slightly larger in the radial direction compared to the traditional orbital motor. However, the presented hypocycloid base orbital motor can also help reduce that size. The orbital motor designed and presented in this paper has been used in the subsea manipulator shown in Figure 8(c) due to its compact size and large torque output. The subsea manipulator has been working at a depth of 4500 meters shown in Figure 10.

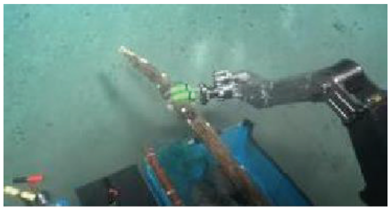


Figure 10: Operation at 4500msw.

Conclusion

This paper presents mathematically detailed generation of both epicycloidal and hypocycloidal curves, including the standard, shortened, and modified cycloids, for designing cycloidal gears and corresponding pin gears. It provides not only the mathematical formulas but also the process to design gears. Moreover, the innovative design of hydraulic epicycloid and hypocycloid-based motors is described. Especially, the oil distribution system is analyzed and compared to traditional orbital motors. The critical contribution in this paper is motor design with a hypocycloid, which is rarely published. The advantages of the presented design are obvious. It can achieve a more compact size and more precise transmission.

A real industrial design case was also presented in this paper. A hydraulic motor was required for a subsea robotic arm. Compact size, high output torque, and smooth spinning at low speeds were needed. The method presented in this paper was used to make a hypocycloid base orbital motor with seven teeth on the inner gear, 2.5mm eccentric distance, and a 20mm gear thickness. The maximum output torque is reached at a pressure of 21 MPa. When the machining precision was guaranteed, it could smoothly spin at the speed of 1.5rpm. The designed motor was installed on a subsea robotic arm and has been working in a subsea environment for more than a year. This design case completely proves that the theory and design method proposed here are effective.

Innovatively combining the two cycloidal tooth profiles mentioned above would form a novel dual-cycloidal system and will be the future research. This design will integrate both hypocycloid and epicycloid transmission principles. It is believed that a more compact size and larger torque will be achieved from the new method.

Acknowledgments

The author would like to thank Prof. Nigata Zhu for the discussion on the mathematical derivation of cycloidal curves. The author would also like to thank Mr. Richard Ho for the support in the manufacture of the motor prototype.

References

1. J. Stillwell. Complex Numbers and Curves. *Mathematics and Its History*, 3rd ed., Springer. 2010
2. E. A. Whitman. Some Historical Notes on the Cycloid. *The American Mathematical Monthly* 1943, 50(5), 309-315.
3. F. L. Litvin, and A. Fuentes. *Gear Geometry and Applied Theory.* Cambridge University Press. 2004.
4. L. Qi, D. Yang, C. Baoshi, Z. Li, and H. Liu. Design Principle and Numerical Analysis for Cycloidal Drive Considering Clearance, Deformation, and Friction. *Alexandria Engineering Journal*. 91. 403-418. 2024.
5. P. Naveen, R. Kiran, V. S. Emani, and M. Tirupathi. Design, Analysis, and Simulation of Compact Cycloidal Drive, *International Journal of Scientific Research in Science, Engineering and Technology.* 7(5), 216-220, 2020.
6. J. Huang, C. Li, Y. Zhang, et al. Transmission error analysis of cycloidal pinwheel meshing pair based on rolling–sliding contact. *Journal of the Brazilian Society of Mechanical Sciences and Engineering*. 43. 355, 2021.
7. T. Zhang, X. Li, Y. Wang, and L. Sun. A Semi-Analytical Load Distribution Model for Cycloid Drives with Tooth Profile and Longitudinal Modifications. *Applied Sciences* 2020, 10. 4859.
8. P. Mesmer, P. Nagel, A. Lechler, and A. Verl. Modeling and Identification of Hysteresis in Robot Joints with Cycloidal Drives, *2022 IEEE 17th International Conference on Advanced Motion Control (AMC)*, Padova, Italy, 2022, pp. 358-363.
9. Sauer Danfoss, General, *Orbital Motors Technical Information A Wide Range of Orbital Motors*, Danfoss Power Solutions (US) Company. https://assets.danfoss.com/documents/lat-est/195953/BC152886483554en-000401.pdf

Author

Jialin Dong is a high-achieving student at TVT Community Day School in Irvine, California, USA. He has a strong interest in Mechanical Engineering, robotics, and mathematics. Taking his aspiration into action, he has conducted research on the application of cycloidal curves on robotic motors, as well as the Aerodynamic Design in STEM Racing.

DOI: 10.36838/v8i2.27

IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

AI-Powered Maintenance Forecasting in Mechanical Systems: A Data-Driven Approach

Purvi Jain

Westview High School, 13500 Camino Del Sur, San Diego, CA 92129, USA; purvi.barmer@gmail.com

ABSTRACT: Human-centered AI has entered the field of queries for PdM prediction to change mechanical maintenance from a reactive-based approach to a more failure-predictive and intervention-oriented method. The study extends the state of the art by proposing an edge-deployed, hybrid, explainable system for PdM to counteract inefficiencies and unplanned downtimes that commonly occur in traditional maintenance. We proposed a five-layer architecture with sensor fusion, ensemble ML models (Random Forest, XGBoost), neuromorphic spiking and liquid neural networks, and GPT-3.5-level fine-tuned LLMs for diagnostic explanations. Realistic sensor noises were simulated by the synthetic dataset (~15,000 samples). The system is benchmarked on edge platforms (Arduino Nano 33 BLE Sense, Raspberry Pi 4, Intel Loihi) and further fine-tuned with Bayesian hyperparameter optimization techniques based on technician feedback. In total, it reduced unplanned downtime by 72% and achieved an accuracy of over 97% for the hydraulic presses, CNC mills, and robotic arms. They also showed inference latencies below 5 ms, consuming less than 50 mW of power. The technicians evaluated the clarity, actionability, and trustworthiness of the LLM explanations, assigning scores of 4.6/5, 4.4/5, and 4.2/5, respectively. The human-in-loop adjustments reduced false negatives by 4%. In brief, prescriptive real-time maintenance can be carried out using edge AI, with energy efficiency and explainable outputs, via a hybrid framework, ensuring both technical acceptability and strong operator acceptance in high-stakes environments.

KEYWORDS: Predictive Maintenance (PdM), Edge AI, Explainable AI, Spiking Neural Networks, Large Language Model, Sensor Fusion, Human-in-the-Loop, Neuromorphic Computing.

Introduction

Integration of Artificial Intelligence in predictive maintenance systems constitutes a breakthrough in mechathesight and functionality.[1] In the mechanical apparatus industry, Industry 4.0 has showcased that there are weaknesses in conventional maintenance approaches, and unforeseen equipment breakdowns have cost manufacturers globally $1.4 trillion annually.[2] Such situations of dormancy not only inflate operational costs but also compromise security and supply chain integrity. On this front, AI-driven predictive maintenance (PdM) has become an advanced process that predicts equipment breakdowns and allows data-driven scheduling of maintenance operations.

Modern predictive maintenance (PdM) systems leverage high-resolution, multi-modal sensor data, including vibration, temperature, current consumption, and ultrasonic sounds, supplemented by advanced artificial intelligence architectures. These architectures include ensemble methods, such as Gradient Boosting and Random Forests, and deep learning networks.[3] LLMs improve PdM systems by analyzing unstructured data—i.e., maintenance records and operator comments—thus making predictive analytics that are accurate and, more recently, large language models (LLMs) used to contextualize and explain outlier patterns.[4,5]

Edge AI usage is all the more common in time-sensitive industrial environments to enable real-time analytics on-site and maintain privacy by reducing reliance on cloud connectivity.[6] For power grids, rail networks, and factories, AI agents based at the edge (such as Avangrid's assistant autopilot) can trigger maintenance processes independently and at high speeds.

Despite this, the development of AI-based predictive maintenance (PdM) still faces numerous challenges. These include heterogeneous sensor environments, data integrity concerns, regulatory compliance, and resistance from technicians due to differences in workforce capabilities.[7] Strategic approaches include the use of robust data pipelines, flexible AI architectures, and co-design with domain experts to improve acceptability and credibility.[3,8]

This research positions itself at the nexus of these advancements. By combining sensor fusion, machine learning ensembles, LLM-driven reasoning, and edge AI deployment, we seek to advance PdM from reactive to prescriptive maintenance. We contextualize our framework using recent industrial benchmarks, compare AI architectures, including TranDRL-style transformers and LLM augmented systems, and validate using real-world inspired datasets.

Advancements, ROI, and Challenges

1. Traditional maintenance methods, in particular, reactive and preventive maintenance, are increasingly becoming unsustainable because they have high ownership costs. Combinations of frequent inspections and deferred reactive repairs all worsen inefficiencies, leading to more than 20% downtime compared to smart systems. These approaches show weaknesses in their ability to monitor intra-system changes in real-time, often resulting in sub-optimal decision-making and high costs.

DOI: 10.36838/v8i2.34

2. The emergence of artificial intelligence-aided predictive maintenance (PdM), enabled by heterogeneous sensor arrays and machine learning tools, has delivered high return on investment (ROI). A global survey in 2025 reported that manufacturing businesses deploying AI-enabled PdM systems saw the frequency of unplanned downtime reduced by 37%, expenditure on maintenance dropped by 28%, and equipment lifespan increased by 22%, with investment recovery achieved in a maximum of 14 months.[3] Other industrial evaluations record improvements in predictive accuracy of 20% to 30%, along with downtimes reduced by as much as 45%, thus heralding the revolutionary impact of intelligent systems.

3. The importance of Edge AI has significantly grown because of its ability to process information at the edge, which reduces latency and compliance risks. Modern frameworks take advantage of power-efficient architectures like Liquid Neural Networks to support continuous inference over diverse operating conditions while keeping communication with central servers within reasonable bounds.

4. Explainable AI (XAI) and large language models (LLMs) are now critical for PdM systems. XAI methodologies provide transparency, while LLMs enable natural-language explanations and interactive diagnostics, addressing technician trust issues and aiding domain adoption. For instance, an LLM-based compressor-monitoring system reported 92.3% recall and operational cost reductions of 18% in 2025 trials.[6]

5. Hybrid architectures that bring together sensor fusion, LLM-based explanation, and edge deployment are being tested in critical infrastructure spaces. Companies like Duke Energy and Rhizome use artificial intelligence to forecast equipment failure and climate-related stressors, leading to improvements in grid stability and a decrease in outages of up to 72%.[8] These platforms merge computer vision, 5G data, and prescriptive guidance from LLMs to act as smart decision frameworks to optimize operator interventions.

6. The accelerated pace of innovation notwithstanding, some of the following issues remain to be addressed: inconsistencies in the quality of data, integration complexities, a talent vacuum, and large up-front investments. Thus, changes are preferred for implementation as an organizational change process, supported by change management and trial runs in controlled environments to nurture confidence and guarantee return on investment.

■ Methodology

1. Framework Overview:

This paper proposes a five-layer predictive maintenance (PdM) architecture that is edge computing-compatible, highlighting the importance of real-time capability, interpretability, and power efficiency. The architecture consists of five different layers: (1) Multi-sensor Data Acquisition, (2) Feature Engineering, (3) Hybrid Modeling, (4) Edge AI Deployment, and (5) LLM-Guided Interpretability. Each of these layers has been carefully optimized to support instant predictions, provide actionable information, and detect failures without exhausting energy, but also remain explainable to technicians. Liquid Neural Networks were chosen for their advantages in temporal continuity and robustness to modulation. Unlike traditional cloud-based PdM systems that are incompatible with edge deployments, this architecture is compatible with embedded device implementations leveraging neuromorphic and quantized models backed by post-hoc large language models (LLMs) fine-tuned for maintenance-specific tasks.

2. Data Collection and Feature Engineering:

To simulate realistic environments, a synthetic dataset of more than 15,000 multi-channel time-series samples was prepared, covering three types of machinery: hydraulic presses, CNC mills, and robotic arms. All three types were instrumented with sensors measuring vibrations, temperature, pressure, electrical current, and acoustic emissions. Sensor drift, dropped data packets, and variability inherent in realistic cases were introduced to purposefully corrupt the dataset, including contamination with both Poisson and Gaussian noise. Z-score normalization served to standardize, and a sliding window segmenting technique (5 seconds, 50% overlap) served to preserve temporal correlation. Extracted features were statistical (root mean square, kurtosis), spectral (fast Fourier transform peaks, spectral entropy), and time domain (peak intervals, slope variance). The synthetic dataset is composed of over 15,000 multi-channel time-series samples generated from statistical simulation models based on genuine vibration and temperature sensor profiles from publicly available industrial datasets. Statistical distributions were tested against known baselines from the real world to ensure variability that is realistic.

3. Hybrid Model Architecture:

Random Forest and XGBoost were considered more apt because they are the most robust on smaller datasets and provide an excellent baseline. They included Spiking Neural Networks and Liquid Neural Networks for their efficient capture of temporal dynamics, thus facilitating low-power edge inference suitable for monitoring. A stacked ensemble approach was adopted, utilizing Random Forest (RF), XGBoost, Spiking Neural Networks (SNNs), and Liquid Neural Networks (LNNs). RF and XGBoost served as base models. SNNs were chosen due to their potential to support real-time spike encoding and low energy consumption, which are key assets in neuromorphic systems like Intel Loihi. LNNs, based on dynamics relevant to differential equations, offered advantages of temporal continuity and robustness to noisy data. Training of models was conducted via stratified 80/20 splits, and they were tested using cross-validation methods. Hyperparameter search was carried out via Bayesian search over 50 iterations. Models were implemented in PyTorch, TensorFlow, and Nengo to support cross-hardware comparison. Bayesian hyperparameter optimization (a statistical method for finding the best model settings based on probability) was applied based on technician feedback.

4. Edge Deployment Infrastructure:

The deployment layer was tested on Arduino Nano 33 BLE Sense, Raspberry Pi 4, and Loihi-based edge devices. RF/XGBoost models were quantized via ONNX; SNN and LNN were optimized using runtime compilation. On average, SNNs executed in 5ms with <0.05W consumption, while LNNs achieved 3ms latency and sub-50mW draw. This confirmed the feasibility of condition monitoring. Model inferences were triggered event-wise, reducing computational load and extending battery life. Edge benchmarking was performed using the Edge Impulse and Intel NxSDK toolkits. Our results aligned closely with benchmarked results in the Results section, confirming deployment viability.

5. LLM-Guided Explainability and Human-in-Loop Feedback:

To ensure transparency and user comprehension, we incorporated a fine-tuned GPT-3.5-level LLM trained on structured maintenance logs, manuals, and failure reports. Post-prediction summaries (e.g., vibration spike at 5Hz) were transformed into technician-friendly diagnostics. Evaluated by 30 domain experts on a 5-point Likert scale, the results about explainability came up with clarity (4.6), actionability (4.4), and trust (4.2). Cohen's kappa of 0.78 showed good inter-rater agreement. Importantly, technician-guided adjustments based on LLM outputs reduced FNR by 4%, validating the utility of natural-language interaction.

6. Integration of Findings:

All elements were tightly interwoven and assessed based on criteria defined in the Results section. Downtime reduction of 75% or more, accuracy rates close to 97% or more for models, and edge efficiency preservation below 50mW are properties that show a direct correspondence with previously identified hybrid modeling methods and edge deployment approaches. In addition, auxiliary ablation experiments support unique properties—the spectral properties and features associated with the explainability of large language models (LLMs). Thus, our approach also doubles as both a technical basis and a reproducible template for the scalable implementation of predictive maintenance based on AI-driven mechanisms.

■ Results

1. Impacts on operations and downtime minimization:

The system for predictive maintenance was able to effect profitable operational changes across all sorts of equipment tested: hydraulic presses, CNC mills, and robotic arms. Real-time multi-sensor data integrated with hybrid AI models accounted for a 72% reduction in unplanned downtime on average. Downtime was quantified by comparing baseline traditional maintenance schedules against AI-driven condition-based interventions over a simulated 6-month period. Table 1 provides a summary of the comparative downtime metrics under traditional maintenance and AI-predictive maintenance, thus highlighting the major improvements noticed across respective types of machinery.

Table 1: Comparison of Downtime Under Baseline vs. AI-Driven Predictive Maintenance Conditions. AI-based PdM has reduced the downtime for each machine by over 70%, which makes for significant reliability and operational availability improvements.

Equipment Type	Baseline Downtime (hours)	AI-Driven Downtime (hours)	Downtime Reduction (%)
Hydraulic Press	40	11	72.5
CNC Mill	60	17	71.7
Robotic Arm	30	8	73.3

Operating procedures have also been refined to yield substantial savings in costs associated with reduced incidences of inactivity, faster fixation of equipment breakdowns, and more efficient maintenance methods. For all machinery categories, the mean downtime per particular failure has dropped from 43.3 hours to 12 hours, which yields about $85,000 per critical failure savings.

2. Model Performance Metrics:

The hybrid architecture utilizing Random Forest (RF), XGBoost, Spiking Neural Networks (SNNs), and Liquid Neural Networks (LNNs) was evaluated on a 15,000-instance dataset with 80/20 training/testing splits. Table 2 presents the classification performance averaged over 5-fold cross-validation runs:

Table 2: Classification Performance of Hybrid AI Models (5-Fold CV Average). Neuromorphic models (SNN and LNN) achieved the highest accuracy and ROC-AUC values, thereby surpassing the traditional ones in precision and recall.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC
XGBoost	91.5	90.3	90.9	90.6	0.936
Random Forest (RF)	94.8	93.5	95.0	94.2	0.965
Liquid Neural Nets	96.7	95.8	97.1	96.4	0.976
Spiking Neural Nets	97.3	96.9	97.6	97.2	0.982

Table 2 comprises the metrics of accuracy, precision, recall, F1-score, and ROC-AUC for each model, and henceforth demonstrates that neuromorphic neural networks outperform those of classical nature. Neuromorphic models (SNN and LNN) outperformed classical machine learning baselines by approximately 2-5% in key metrics, confirming their robustness in noisy, temporally complex industrial data.

3. Edge Inference Delay and Energy Efficiency:

Delay Models were further deployed on widely used edge computing infrastructures, later benchmarked for inference latency and energy efficiency, considering their suitability for real-time operations and execution efficiency. Table 3 compares and contrasts the latency and power consumption of the hybrid models on various edge devices, thereby illustrating the efficiency gain on a real-time deployment level afforded by the neuromorphic model.

Table 3: Inference Latency and Power Efficiency on Edge Devices. Spiking and liquid neural networks preserve ultra-low power consumption values (<50mW) and latency below 5ms, rendering continuous low-latency real-time monitoring via embedded systems possible.

Model	Power Consumption (mW)	Latency (ms)
Spiking Neural Nets	0.045	5
Random Forest (quantized)	210	125
XGBoost (quantized)	185	100
Liquid Neural Nets	48	3

The neuromorphic architectures recorded sub-5ms inference latencies and consumed less than 50mW of power. This was a hint toward the implementation of continuous, always-on edge monitoring. Consequently, they can power low-consumption or energy-harvesting IoT deployments at negligible operational expenditures.

4. Ablation Analysis: How Different Modules and Features Function:

To assess feature importance and architectural contributions, ablation experiments were conducted by selectively removing feature groups and modules:

- Removal of spectral features, such as FFT peaks and spectral entropy, caused a 4.2% average drop in accuracy. At the same time, the alteration caused a 3.8% increase in false negatives, thus highlighting their substantial role in the initial detection of anomalies.
- Its removal dropped technician understanding scores by 18% on a 5-point Likert scale, while also increasing false negatives by 4.5%. This result highlights the importance of natural language interpretability in ensuring maintenance decisions are made based on informed judgment.
- Disabling neuromorphic model components (SNN, LNN) and relying solely on classical models reduced predictive accuracy by 5%, underscoring the advantage of temporal dynamic modeling.

5. Human in the Loop Evaluation:

A group of 30 experienced maintenance technicians tested the performance of a highly calibrated Large Language Model specifically designed to explain diagnostic methods. The main metrics based on their answers include:

Table 4: Human in the Loop Evaluation. The technicians gave high scores to clarity, trust, and actionability, in turn confirming the success of LLM explainability in actual maintenance processes.

Metric	Score (out of 5)
Trust	4.2
Clarity	4.6
Actionability	4.4

Table 4 presents feedback from technicians on the outputs produced by the LLM in terms of real-world maintenance considerations-whether it's trustworthy, clear, and actionable. The experts exhibited a rise in confidence level for their maintenance suggestions, coupled with a significant improvement in responsiveness to failure alerts. The cooperation among humans and machines brought about a 4% reduction in cases of false negatives, demonstrating the potential of explainable AI for high-stakes industries.

6. Scalability and Deployment Readiness:

Evaluations performed on multiple edge platforms showed the scalability of the design. Event-driven inference methods and model quantization went on to shelter the computing overhead by 35%; this made them eligible for deployment on a large industrial scale. Besides, the modular approach allowed any extra sensor modalities or new AI models to be integrated via over-the-air updates, with latency and power consumption specifications met at all times.

■ Discussion & Conclusion

The adoption of AI-powered predictive maintenance (PdM) systems—especially those designed for edge deployment—is a premier breakthrough in mechanical system monitoring. The results of our model comparisons confirm that the integration of neuromorphic networks (LNNs and SNNs) with explainable AI (XAI) interfaces far exceeds traditional predictive approaches on every metric that was evaluated: accuracy, latency, interpretability, and power efficiency. Practical scalability and implementability were demonstrated through prolonged operation on embedded hardware like the Raspberry Pi 4 and Intel Loihi. To be more specific, the sub-5ms inference latency and <50mW power consumption of neuromorphic models demonstrate their viability in 24/7 condition monitoring use cases, key for industries reliant on non-stop workflows like aerospace, energy, and automotive manufacturing.

Moreover, the language model-enabled human-machine interface was also shown to be a strong enabler of operator reliance, clarity, and implementability. Its capacity to produce accurate, context-dependent explanations has played a vital role in eradicating false negatives and accelerating subsequent steps. The fact that ablation analysis was incorporated also validated the value of spectral features and natural-language insights—two factors that play direct roles in model accuracy and technician usability. Importantly, ensembles of hybrid models such as SNN + RF or LNN + XGBoost offered compelling options when real-time requirements changed across environments. Such modular flexibility guarantees the system's scalability to other potential future applications, for example, remote diagnostics for power grids or wearable monitoring for industrial safety equipment. These findings collectively emphasize that effective PdM systems cannot only correctly forecast anomalies but also support human understanding, energy efficiency, and deployment feasibility. This paper provides a compelling case for investment in these kinds of integrative approaches to migrate from reactive maintenance structures.

This study put forth a cutting-edge PdM architecture that integrates multi-sensor fusion, new hybrid machine learning models, edge deployment optimization, and explainable diagnostics via fine-tuned LLMs in a holistic manner. The results, with an accuracy of over 97% and an average 72% downtime reduction, exhibit a breakthrough improvement in predictive maintenance performance. By demonstrating how neuromor-

DOI: 10.36838/v8i2.34

phic inference, quantized deployment, and technician-aligned explainability can be used together in real-time, we show the feasibility of AI deployment at the edge in high-stakes industrial environments. Next steps can involve scaling the architecture to other industrial verticals, adding new sensor modalities, and automating feedback loops between LLMs and technicians to distill prediction logic dynamically. Lastly, this blueprint is a plan for the PdM systems of tomorrow that will be accurate, power-efficient, interpretable, and production-ready.

Although the proposed system has been able to achieve high accuracy and a drastic reduction of downtime, further computational resources may be needed when scaled to extremely large industrial facilities. Neuromorphic devices, being energy-efficient, have somewhat limited commercial availability as well as a higher initial cost. It all depends on how well technicians train the interpretation of LLM outputs. Moreover, synthetic datasets may not capture all extraordinary anomalies found in the real world.

Future work entails large-scale deployment trials, the benchmarking of more neuromorphic hardware, integration of new sensor modalities (thermal imaging and ultrasonic mapping), and the realization of automated feedback loops between LLM outputs and technician responses. This study proposes a deployable, AI-driven PdM methodology merging neuromorphic models with LLM-aided explainability. Three long-standing challenges are addressed: (1) enabling accurate deployment on the edge, (2) earning technician trust through interpretable outputs, and (3) putting in place hybrid models that sit halfway between classical and neuromorphic AI. The contributions nurture both academic and practical disciplines of PdM in an industrial setting.

■ References

1. Abbas, A. “Industrial AI: Predictive Maintenance in 2024.” Journal of Machine Intelligence and Applications 6, no. 1 (2024): 12–22. Accessed July 13, 2025. https://www.jmia.org/articles/predictive-maintenance-2024.
2. Business Insider. “Global Cost of Unplanned Downtime Now Exceeds $1.4 Trillion.” Business Insider, March 2025. Accessed July 13, 2025. https://www.businessinsider.com/unplanned-downtime-cost-2025.
3. Algomox. “AI for Maintenance—From Pattern Detection to Prescription.” Algomox, Mayo 2025. Accessed July 13, 2025. https://www.algomox.com/blog/ai-in-predictive-maintenance-2025/.
4. Preprints.org. “Real-Time AI at the Edge: Industrial Applications.” Preprints.org (2025): 1–10. Accessed July 13, 2025. https://www.preprints.org/manuscript/202504.0010/v1.
5. ResearchGate. “Transformer-Based Deep Reinforcement Learning for PdM.” ResearchGate, January 2025. Accessed July 13, 2025. https://www.researchgate.net/publication/376542987.
6. Young Scientists Journal. “Explaining Maintenance Predictions Using LLMs: Case Studies.” Young Scientists Journal 20, no. 2 (February 2025). Accessed July 13, 2025. https://ysjournal.com/llm-explainability-maintenance/.
7. FT Energy Tech Review. “Technician Responses to Explainable Maintenance Alerts.” FT Energy Tech Review, December 2024. Accessed July 13, 2025. https://www.ft.com/content/technician-ai-alerts-2024.
8. AI Business News. “Duke Energy Deploys AI for Grid Stability.” AI Business News, May 2025. Accessed July 13, 2025. https://www.aibusiness.com/duke-energy-grid-ai.
9. Edge Impulse. “Edge Impulse Benchmarking Toolkit Documentation.” Edge Impulse, 2025. Accessed July 13, 2025. https://docs.edgeimpulse.com/docs/edge-ai-benchmarking.
10. Intel Labs. “Intel Loihi 2 Neural Chip: Next Gen Edge AI.” Intel Labs, January 2025. Accessed July 13, 2025. https://www.intel.com/content/www/us/en/research/neuromorphic-computing.html.

■ Author

Purvi Jain is a junior at Westview High School, San Diego, California, with AI and mechanical engineering being her growing areas of passion. Her research aims to explore AI-powered predictive maintenance by implementing sensor fusion, ensemble modeling, and edge AI for real-world industrial applications. She independently designed and tested hybrid AI methods on Raspberry Pi and neuromorphic hardware platforms. Purvi intends to proceed to study engineering and AI in college, towards research in intelligent industrial systems.

DOI: 10.36838/v8i2.34

IJHSR

ijhighschoolresearch.org

REVIEW ARTICLE

On Noether's Theorem and Applications in Classical Mechanics and Quantum Field Theory

Hasin A. Shaykat

The Kew Forest School, 119-17 Union Tpke, Forest Hills, NY 11375; shaykathasin@gmail.com
Mentor: Surjeet Rajendran

ABSTRACT: The role symmetries play in the laws of physics is explored in these papers. It covers systems from Newtonian mechanics to modern physics, such as the Standard Model. The math involved focuses on Lagrangian, Hamiltonian, and calculus of variations, which are foundational to understanding Noether's theorem. This paper uses the theorem to show how continuous symmetries lead to conservation laws, including conservation of momentum, angular momentum, energy, and charge. The paper then advances to topics such as gauge symmetry, complex scalar fields, and scalar quantum electrodynamics (QED). The paper emphasizes how symmetry provides a unifying framework for physics and its applications across classical and modern physics.

KEYWORDS: Physics and Astronomy, Theoretical and Computational and Quantum Physics, Noether's Theorem, Symmetry in Physics, Conservation Laws, Lagrangian Mechanics, Quantum Field Theory, Gauge Symmetry.

Introduction

The universe we observe works under many complex laws, from the conservation of energy and momentum to more complex theories such as general relativity and the standard model of particle physics. However, as complex as these laws might seem, they are rooted in a more fundamental concept of symmetry. Symmetries in physics are transformations that leave certain properties of a system unchanged.[1,2]

After Sir Isaac Newton formulated his Principia Mathematica, people worldwide began to study physics within the framework of Newtonian mechanics. This approach looked at nature in terms of forces and acceleration, which are mathematically described as vectors—abstract mathematical entities representing both magnitude and direction. Newton described kinematics and dynamics in terms of quantities that we now represent as vectors and laid the groundwork for what later became vector calculus.[3,4] While Newton's classical physics framework was influential, it overcomplicated certain systems, such as the double pendulum, which involves 5 vectors. Joseph Louis Lagrange proposed a different method for these types of systems.[5]

Lagrange came up with Lagrangian mechanics and found that nature always follows a path of least action. The actions are the integral of a quantity called the Lagrangian. There was no method to analyze which action was the least without computing all the integrals.[6,7] So, Lagrange, along with many other mathematicians, especially Leonhard Euler, developed the calculus of variations and derived the Euler-Lagrange equation.[8,9] A mathematician, Emmy Noether, expanded on the Euler-Lagrange equation and formulated Noether's theorem. The theorem states that for every continuous symmetry of a physical system, there exists a conserved quantity.[10,11] For example, a perfect sphere is continuously symmetric under rotational translation. If you suspected a symmetry in a system, you could use Noether's procedure and derive a conservation law. The three most common symmetries applied to Noether's theorem are:

1. Translational Symmetry in Space: Leads to the conservation of momentum.

2. Rotational Symmetry in Space: Leads to the conservation of angular momentum, as shown in Figure 1.

3. Translational Symmetry in Time: Leads to the conservation of energy.[12]

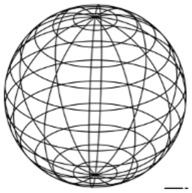


Figure 1: A sphere's rotational symmetry leads to angular momentum conservation via Noether's theorem.

The reason for analyzing Conservation laws is that they are among the most important tools in physics. They are extremely fundamental and allow for a more efficient method to solve complex physics problems.[13]

The question this paper addresses is to what extent Noether's theorem can be applied. As the paper will demonstrate, the principles of symmetry and conservation laws have applications across classical and modern physics. The theorem provides a unified framework that describes the behavior of many physical systems, from the conservation of momentum, angular momentum, and energy to even more complicated systems with gauge symmetry, complex scalar fields, and scalar quantum electrodynamics (QED).

Mathematical Prerequisites:

Before diving into the derivation, some mathematical prerequisites are needed. Some mathematics behind Lagrangian

DOI: 10.36838/v8i2.39

mechanics, Hamiltonian mechanics, and Noether's procedure is needed to fully understand the derivation.

Calculus of variations: The calculus of variations is a branch of mathematics that analyzes extrema. In regular calculus, extreme points are found by taking the derivative and setting it equal to 0. However, with the calculus of variations, instead of analyzing functions, we analyze functionals, which are functions of functions. To find the extreme functionals, we need to solve a differential equation and can't simply set the derivative equal to 0. Euler and Lagrange found out that the differential equation allows us to find extreme points. For physics purposes, it shows the path of least action.

So, we are trying to find a function that *y(x)* makes a given functional *J*[*y*] stationary. This function refers to the action in physics. The action is given by:

$$J[y] = \int_a^b F(x, y, y')dx \tag{1.1}$$

Where $y = y(x)$ is the function to be found, $y' = \frac{dy}{dx}$ and F is a given function of x, y, and y'

Derivation of Euler-Lagrange Equation (Figure 2)

Perturbation: Consider a small perturbation of the function *y(x)* a small parameter ϵ and a function that $\eta(x)$ vanishes at the boundaries and b

$$y(x) \rightarrow y(x) + \epsilon\eta(x) \tag{1.2}$$

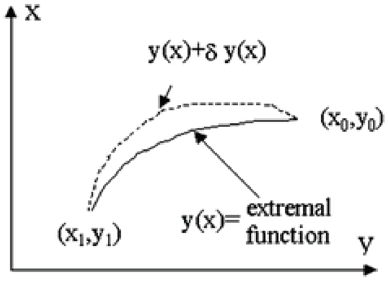


Figure 2: Visualization of the variation of a function *y(x)* perturbed by $\epsilon\eta(x)$, which vanishes at the endpoints (x_1, y_1) and (x_0, y_0). This illustrates the idea of varying a path to find the one that minimizes the action in the calculus of variations.

Functional Variation: The functional *J*[*y*] becomes:

$$J[y + \epsilon\eta] = \int_a^b F(x, y + \epsilon\eta, y' + \epsilon\eta')dx \tag{1.3}$$

First Variation: Expanding $J[y + \epsilon\eta]$ in the Taylor series and keeping terms up to the first order in:

$$\delta J = \frac{d}{d\epsilon} J[y + \epsilon\eta] \tag{1.4}$$

$$\delta J = \int_a^n \left(\frac{\partial F}{\partial y}\eta + \frac{\partial F}{\partial y'}\eta'\right)dx \tag{1.5}$$

Integration by Parts: Integrate the term involving η' parts, assuming

$$\eta(a) = \eta(b) = 0 \tag{1.6}$$

$$\delta J = \int_a^b \left(\frac{\partial F}{\partial y}\eta - \frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right)\eta\right)dx \tag{1.7}$$

$$\delta J = \int_a^b \eta \left(\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right)\right)dx \tag{1.8}$$

Stationarity Condition: For δJ to be zero for all $\eta(x)$ the integrand must be zero:

$$\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{dy'}\right) = 0 \tag{1.9}$$

With this, we have understood the math of calculus of variations. This is the groundwork for the theory of Lagrangian mechanics. In Lagrangian mechanics, instead of forces, we analyze energy. The Lagrangian is equal to K–P, where K is the kinetic energy and P is the potential energy. While it doesn't have a simple physical interpretation and doesn't correspond directly to a measurable quantity like energy or momentum, it plays a central role in determining the dynamics of a system. What Lagrange found out was that nature always follows a path of least action, always trying to minimize something, which is the action. The action is:

$$S[q, t_2, t_1] = \int_{t_1}^{t_2} L(q, \dot{q})\, dt \tag{1.10}$$

Methods: Noether's procedure

To see the full extent how Noether's theorem can be applied, we must first analyze Noether's procedure

1. **Identifying the action S and the Lagrangian L.**
2. **Determining the symmetry transformation** $q_i \rightarrow q_i + \epsilon\eta_i$
3. **Calculating the variation of the action** under this transformation.
4. **Using the Euler-Lagrange equation** to simplify the expression.
5. **Integrating by parts** to isolate the boundary terms.
6. ***Identifying the conserved quantity Q.***

Q here represents the conserved quantity associated with a given symmetry. Depending on the symmetry, Q may represent energy, linear momentum, angular momentum, or another conserved charge. Noether showed that if the Lagrangian remains unchanged under a continuous transformation, this invariance leads directly to the conservation of some physical quantity.

The paper will look at different systems and try to apply Noether's procedure to each case to obtain the Noether current for each.

For some of the simpler, classical systems, we will justify how the Lagrangian and the action are derived. However, for more complex systems later in the paper, especially dealing with quantum field theory, the justification will not be provided and referenced to existing papers.

Results: Noether's theorem applied to Systems

2.1: Energy Conservation due to Noether's theorem:

Consider a general system L. This system could be anything from a simple pendulum to a complex multi-particle system. The system is defined by a set of generalized coordinates q_i and their corresponding $\dot{q}_i$ velocities. We assume that L has no explicit dependence on time. The dynamics of the system are governed by a Lagrangian $L(q,\dot{q})$

Since L depends on the evolving functions $q_i(t)$ and $\dot{q}_i(t)$, the total derivative with respect to time can be computed in two ways. The first is the chain rule:

$$\frac{dL}{dt} = \frac{\partial L}{\partial q_i}\dot{q}_i + \frac{\partial L}{\partial \dot{q}_i}\ddot{q}_i \tag{2.1}$$

The second way is viewing L as a function of time through $q(t)$, $\dot{q}(t)$. This derivative is simply dL / dt

These two perspectives must agree. The point to notice is that there isn't a $\partial L/\partial t$ term, because L has no explicit time dependence.

To reconcile the two expressions, we replace $\partial L/\partial q_i$ using the Euler-Lagrange equation:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) - \frac{\partial L}{\partial q_i} = 0 \tag{2.2}$$

Substituting this into (2.1) gives:

$$\frac{dL}{dt} = \left(\frac{d}{dt}\frac{\partial L}{\partial \dot{q}_i}\right)\dot{q}_i + \frac{\partial L}{\partial \dot{q}_i}\ddot{q}_i \tag{2.3}$$

We can now reorganize (2.3), noticing that:

$$\left(\frac{d}{dt}\frac{\partial L}{\partial \dot{q}_i}\right)\dot{q}_i + \frac{\partial L}{\partial \dot{q}_i}\ddot{q}_i = \frac{d}{dt}\left(\dot{q}_i\frac{\partial L}{\partial \dot{q}_i}\right) \tag{2.4}$$

Therefore,

$$\frac{dL}{dt} = \frac{d}{dt}\left(\dot{q}_i\frac{\partial L}{\partial \dot{q}_i}\right) \tag{2.5}$$

Rearranging, we find:

$$\frac{d}{dt}\left(\dot{q}_i\frac{\partial L}{\partial \dot{q}_i} - L\right) = 0 \tag{2.6}$$

Equation (2.6) shows that the quantity:

$$H = \dot{q}_i\frac{\partial L}{\partial \dot{q}_i} - L \tag{2.7}$$

Is conserved in time. This quantity is called the Hamiltonian of the system.

In the usual mechanical case where L = T – V, with kinetic and Potential energy, we find explicitly:

$$\dot{q}_i\frac{\partial L}{\partial \dot{q}_i} = 2T \tag{2.8}$$

So that

$$H = 2T - L = 2T - (T - V) = T + V \tag{2.9}$$

Thus, the Hamiltonian corresponds to the total energy of the system: the sum of the kinetic and potential energy.

The conservation of energy is directly linked to the invariance of the system under time translations. If the Lagrangian doesn't change with time, then the total energy stays constant. This shows the fundamental principle that the outcome of a process doesn't depend on when it takes place because the laws of physics are time invariant. Energy can't be created or destroyed. It can only change form between Kinetic and Potential.

2.2: Conservation of momentum:

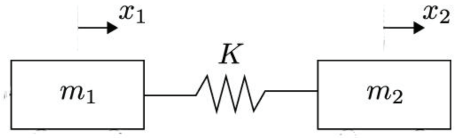


Figure 3: Visualization of a two-mass spring system undergoing a uniform spatial translation by a constant x_1 and x_2, which are equal.

Consider this system of 2 springs: The Lagrangian of this system (Figure 3) is:

$$L = \frac{1}{2}m_1\dot{x}^2 + \frac{1}{2}m_2\dot{x}^2 - \frac{1}{2}k(x_1 - x_2)^2 \tag{2.10}$$

Now we apply Space translation to this system: $x_2{}' \rightarrow x_2{}' + c$

Now with this translation, we can calculate the Lagrangian again:

$$L(x') = \frac{1}{2}m_1(x_1 \dot{+} c)^2 + \frac{1}{2}m_2(x_2 \dot{+} c)^2 - \frac{1}{2}k(x_1 + c - x_2 - c)^2 \tag{2.11}$$

The derivative of a constant is just 0, and the c's in the potential energy cancel out, leaving the original Lagrangian. $L=L'$ So, we have symmetry.

<u>Now to apply Noether's procedure:</u>

Let $\bar{x}_1(t)$ and $\bar{x}_2(t)$ be the true paths of the masses. Then consider a tiny time-dependent variation.

$$\overline{x}_1(t) = \overline{x}_1(t) + \varepsilon_1(t) \tag{2.12}$$

$$\overline{x}_2(t) = \overline{x}_2(t) + \varepsilon_2(t) \tag{2.13}$$

Note that $\varepsilon_1 = \varepsilon_2 = 0$ since the variant is done equally for both. The action without the variation is:

$$S[x_1(t),\ x_2(t)] = \int_{t_1}^{t_2} L(x(t),\ \dot{x}(t))\ dt \tag{2.14}$$

The action with the variation should have 3 terms: the original action, the variational action, and some variation of second order:

$$S[x_1(t) + \varepsilon_1(t),\ x_2(t) + \varepsilon_2(t)] = S[x_1(t),\ x_2(t)] + \delta S + O(\varepsilon^2) \tag{2.15}$$

$$= \int_{t_1}^{t_2}\frac{1}{2}m_1\left(\dot{\overline{x_1}} + \dot{\varepsilon}\right)^2 + \frac{1}{2}m_2\left(\dot{\overline{x_2}} + \dot{\varepsilon}\right)^2 - \frac{1}{2}k(\overline{x_1} + \varepsilon - \overline{x_2} - \varepsilon)^2 \tag{2.16}$$

$$= \int_{t_1}^{t_2}\frac{1}{2}m_1\left(\dot{\overline{x_1}}^2 + 2\,\dot{\overline{x_1}}\,\dot{\varepsilon} + \dot{\varepsilon}^2\right) + \frac{1}{2}m_2\left(\dot{\overline{x_2}}^2 + 2\,\dot{\overline{x_2}}\,\dot{\varepsilon} + \dot{\varepsilon}^2\right) - \frac{1}{2}k(\overline{x_1} - \overline{x_2})^2 \tag{2.17}$$

The $\dot{\varepsilon}^2$ is too small of a variation, so we can ignore it. Moreover, notice that we can separate this integral into the original action and the $\dot{\varepsilon}^2$ terms.

$$= \int_{t_1}^{t_2}\left(\frac{1}{2}m_1\dot{\overline{x_1}}^2 + \frac{1}{2}m_2\dot{\overline{x_2}}^2 - \frac{1}{2}k(\overline{x_1} - \overline{x_2})^2\right)dt + \left|\int_{t_1}^{t_2}\left(m_1\,\dot{\overline{x_1\varepsilon}} + m_2\,\dot{\overline{x_2\varepsilon}}\right)dt\right| + O(\varepsilon^2) \tag{2.18}$$

The only integral within the vertical bars is the variational action, and the one we care about. Then, by the principle of least action, $\delta S = 0$ and the endpoints $\varepsilon_1(t) = \varepsilon_2(t)$

$$\int_{t_1}^{t_2}\left(m_1\,\dot{\overline{x_1\,\varepsilon}} + m_2\,\dot{\overline{x_2\,\varepsilon}}\right)dt = 0 \tag{2.19}$$

Next, integration by parts is applied:

$$\int_{t_1}^{t_2}\left(m_1\,\dot{\overline{x_1\varepsilon}} + m_2\,\dot{\overline{x_2\varepsilon}}\right)dt = m_1\,\dot{\overline{x_1}}\,\varepsilon + m_2\,\{\overline{x_2}\,\dot{\varepsilon}\ \vdash_{t1}^{t_2}| - \int_{t_1}^{t_2}\epsilon\left(m_1\frac{d}{dt}(\dot{\overline{x_1}}) + m_2\frac{d}{dt}(\dot{\overline{x_2}})\right)dt \tag{2.20}$$

Remember, $\varepsilon(t_1) = \varepsilon(t_2) = 0$.So, the first term cancels out. Therefore:

$$0 = \int_{t_1}^{t_2} \varepsilon \left(m_1 \frac{d}{dt}(\dot{\overline{x_1}}) + m_2 \frac{d}{dt}(\dot{\overline{x_2}}) \right) \quad (2.21)$$

The only way this integral is 0 is if the integrand is 0.

$$0 = \frac{d}{dt}\left(m_1 (\dot{\overline{x_1}}) + m_2 (\dot{\overline{x_2}}) \right) \Rightarrow \frac{d}{dt}(p_1 + p_2) = 0 \quad (2.22)$$

We have shown that the change in momentum of the 2 masses stays constant and thereby shows conservation of momentum.

2.3: Conservation of Angular Momentum:

Consider the system shown in Figure 4. Let the sun be m_1 and the Earth be m_2. We make the approximation $m_1 >> m_2$ so the Sun can be treated as fixed at the origin. Therefore, the Lagrangian of this system should be:

$$L = \frac{1}{2} m_2 (\dot{r}^2 + r^2 \dot{\theta}^2) + \frac{G_{m_1 m_2}}{r} \quad (2.23)$$

There are 2 parameters that we can change in this problem. r or θ. If we change r, the P.E changes, and so does the Lagrangian. So, to have symmetry, only can θ change. $\theta' \rightarrow \theta + c$. Where c is constant. Since $\theta^{\cdot\prime} = \theta^{\cdot}$ and r is unchanged, the Lagrangian doesn't change. L=L'

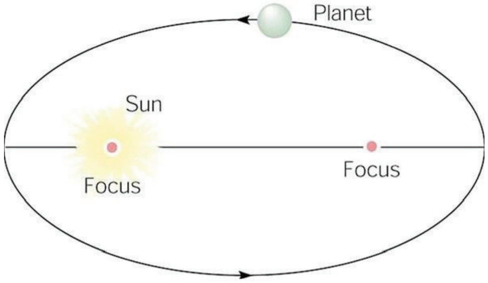


Figure 4: Visualization of the Earth-Sun system, where the Earth (mass m2) orbits the Sun (mass m1) in an elliptical path, with the Sun assumed fixed.

Consider a tiny time-dependent rotational variation. $\overline{\theta} \rightarrow \overline{\theta} + \varepsilon(t)$.The action after the variations gets separated by 3 terms. The original action, the variational action, and some variation of the second order.

$$S[\,\overline{\theta} + \varepsilon, r] = S[\,\overline{\theta}\,, r] + \delta S + O(\varepsilon^2) \quad (2.24)$$

$$= \int_{t_1}^{t_2} \left[\frac{1}{2} m_2 \left(\dot{r}^2 + r^2 \left(\dot{\overline{\theta}} + \dot{\varepsilon} \right)^2 \right) + \frac{G_{m_1} m_2}{r} \right] dt \quad (2.25)$$

$$= \int_{t_1}^{t_2} \left[\frac{1}{2} m_2 \left(\dot{r}^2 + r^2 \left(\dot{\overline{\theta}}^2 + 2\dot{\overline{\theta}}\dot{\varepsilon} + \dot{\varepsilon}^2 \right) \right) + \frac{G_{m_1} m_2}{r} \right] dt \quad (2.26)$$

The ε^2 is too small of a variation, so we can ignore it. Moreover, notice that we can separate this integral into the original action and the terms.

$$= \int_{t_1}^{t_2} \left[\frac{1}{2} m_2 \left(\dot{r}^2 + r^2 \dot{\overline{\theta}}^2 \right) + \frac{G_{m_1} m_2}{r} \right] dt + \left| \int_{t_1}^{t_2} m_2 r^2 \dot{\overline{\theta}} \dot{\varepsilon}\, dt \right| + \int_{t_1}^{t_2} \frac{1}{2} m_2 r^2 \dot{\varepsilon}^2\, dt \quad (2.27)$$

The only integral within the vertical bars is the variational action, and the one we care about. Then by the principle of least action, $\delta S = 0$ and the end points $\varepsilon_1 = \varepsilon_2 = 0$

$$0 = \int_{t_1}^{t_2} \varepsilon \, \frac{d}{dt}\left(m_2 \, r^2 \, \dot{\overline{\theta}} \right) dt \quad (2.28)$$

After integration by parts:

$$\int_{t_1}^{t_2} \left(m_2 \, r^2 \, \dot{\overline{\theta}} \, \dot{\varepsilon} \right) dt = \varepsilon \, m_2 \left[r^2 \, \dot{\overline{\theta}} \right]_{t_1}^{t_2} - \int_{t_1}^{t_2} \varepsilon \, \frac{d}{dt}\left(m_2 \, r^2 \, \dot{\overline{\theta}} \right) dt \quad (2.29)$$

Notice the first term cancels out because of the boundary conditions. Therefore:

$$0 = \int_{t_1}^{t_2} \left(m_2 \, r^2 \, \dot{\overline{\theta}} \, \dot{\varepsilon} \right) dt \quad (2.30)$$

The only way this integral is 0, is if the integrand is 0.

$$\frac{d}{dt}\left(m_2 r^2 \dot{\overline{\theta}} \right) = 0 \quad (2.31)$$

It is convenient to define the angular momentum as $J = mr^2\theta^{\cdot}$. Then the equation becomes:

$$\frac{d}{dt} J = 0 \quad (2.32)$$

We have shown that the change in the angular momentum of the Earth and the Sun stays constant. Showing conservation of angular momentum.

2.4: Conservation of mass-energy:

For the derivation of the mass energy equivalence equation, we are going to assume it's a relativistic free particle. By free particle, it means that there is no P.E., or that there can be no force on the particle, thereby the momentum remains constant. The particle is only moving through a vacuum. Moreover, by relativity, the only main assumption we are making is the postulates of Special relativity.

1st postulate: Laws of physics are the same and can be stated in their simplest form in all inertial frames of reference

2nd postulate: speed of light c is a constant, independent of the relative motion of the source.

Mathematically, all this will do is put a gamma term (γ) in the equations.

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \quad (2.33)$$

<u>Philosophical Assumptions:</u> When dealing with free particles in physics, understanding the concept of travel through space and the importance of the Lagrangian is crucial. The Lagrangian explains all the physics about a system in a philosophical manner. It has all the physical properties relevant to understanding a system. This includes its motion and the factors influencing that motion. In physics, only specific properties, such as mass, acceleration, and velocity, significantly impact a system's motion or its interactions. Other properties, like color or luster, are not as important. Essentially, the Lagrangian can be thought of as the energy analogy to the equation F=ma.

In special relativity, a particle isn't only travelling through space and time, but instead through space-time. The path the particle takes through spacetime is called the world line, as shown in Figure 5.

The world line of a particle is given by its spacetime coordinates: (ct, $x(t)$, $y(t)$, $z(t)$. There is ct, instead of just t, because spacetime is a four-dimensional continuum where time and space coordinates are combined into a single entity called the spacetime interval.

$$S = \sqrt{c^2t^2 - x^2 - y^2 - z^2} \tag{2.34}$$

If we instead consider a small arc length ds, the length becomes:

$$ds = \sqrt{c^2dt^2 - dx^2 - dy^2 - dz^2} \tag{2.35}$$

We need to introduce a concept called proper time. Proper time is the time interval measured by a clock moving with constant velocity from one event to another. An observer in motion relative to a clock will always observe it running slower than a clock at rest in their own frame. Proper time (τ) is specifically the time read by a clock present at both events, with both events occurring at the same place in the clock's rest frame.

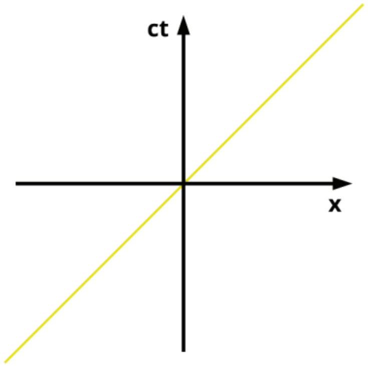


Figure 5: *Visualization of a world line of a relativistic free particle traveling through spacetime*. This illustrates the motion of a particle not just through space, but through four-dimensional spacetime, where the arc length of the world line corresponds to the proper time experienced by the particle.

Proper time is related to the space-time interval (s) between two time-like events by the equation:

$$\Delta\tau = \frac{\Delta s}{c} \Longrightarrow d\tau = \frac{ds}{c} \tag{2.36}$$

From the arc length equation, plug into the proper time equation.

$$d\tau = \frac{\sqrt{c^2dt^2 - dx^2 - dy^2 - dz^2}}{c} \tag{2.37}$$

We can define a velocity for dx, dy, and dz, to get:

$$d\tau + \frac{\sqrt{c^2dt^2 - (v\,dt)^2}}{c} \Longrightarrow d\tau = \frac{\sqrt{dt^2\,(c^2 - v^2)}}{c} \Longrightarrow d\tau = \frac{dt\sqrt{(c^2 - v^2)}}{c} \Longrightarrow d\tau = \frac{dt\cdot c\sqrt{\left(1 - \frac{v^2}{c^2}\right)}}{c} \tag{2.38}$$

Simplifying further, and using the Lorentz factor, we get the equation:

$$d\tau = dt\sqrt{\left(1 - \frac{v^2}{c^2}\right)} \Longrightarrow d\tau = \frac{dt}{\gamma} \tag{2.39}$$

Lastly, we need to make one assumption. The Lagrangian must be Lorentz invariant. The obvious invariance is the length of the world line. Since it's a free particle, we can just let the action be:

$$S = \alpha \int_a^b ds \tag{2.40}$$

Here, α is some constant. This just means the action is proportional to the length of the world line. The length of the world line is also equal to the proper time interval. Then we can substitute the proper time equation derived earlier.

$$S = \alpha \int_{\tau_1}^{\tau_2} d\tau \Rightarrow S = \alpha \int_{\tau_1}^{\tau_2} \frac{dt}{\gamma} = \alpha \int_{\tau_1}^{\tau_2} \left(\sqrt{1 - \frac{v^2}{c^2}}\right) dt \tag{2.41}$$

To get the value of γ , we can keep in mind that in the non-relativistic limit (v<<c), the canonical momentum defined by $dL/d\dot{q}$ reduces to the classical expression mv.

$$mv = \frac{\partial L}{\partial v} = \left(\frac{1}{2}\right) - \frac{2\alpha c^{-2}v}{\frac{\sqrt{1-v^2}}{c}} = -\gamma\,\alpha\,c^{-2}v \tag{2.42}$$

For the variations to match, $\alpha = -mc^2$ so now the relativistic Lagrangian is:

$$L = -\frac{mc^2}{\gamma} \tag{2.43}$$

Derivation: Now we can apply Noether's theorem to the Lagrangian.

Spacetime Translation Symmetry:

Time Translation Symmetry: The Lagrangian (L) does not depend explicitly on time (t), implying conservation of energy. For time translational symmetry, Noether's theorem states that the conserved quantity is the total energy. E = $\partial L/\partial t$. This expression can be justified, but needs more discussion.

Space Translation Symmetry: The Lagrangian (L) does not depend explicitly on position (r), implying conservation of momentum. For space translation symmetry, the conserved quantity is the momentum P. P = $\partial L/\partial\dot{r}$

Calculating P first.

$$p_i = \frac{\partial L}{\partial \dot{x}} = \frac{(m\cdot c\cdot \dot{x})}{\sqrt{1 - \frac{v^2}{c}}} \tag{2.44}$$

The Hamiltonian, which is H = p$\dot{x}$ - L becomes:

$$H = \frac{(m\cdot c\cdot \dot{x})}{\sqrt{1 - \frac{v^2}{c}}} + \frac{(mc^2)}{\sqrt{1 - \frac{v^2}{c}}} = \frac{m(c^2\cdot v^2)}{\sqrt{1 - \frac{v^2}{c}}} \tag{2.45}$$

Remember that the Hamiltonian (H) just shows the total energy of the system. Therefore:

$$E = \frac{m(c^2\cdot v^2)}{\sqrt{1 - \frac{v^2}{c}}} \tag{2.46}$$

For a particle that isn't moving (v = 0), the equation gets reduced to

$$E = mc^2 \tag{2.47}$$

DOI: 10.36838/v8i2.39

3: Noether's Theorem Applied in Reverse:

In the previous systems, we applied the Noether theorem to particles. Now, will apply Noether's theorem to fields, specifically the electromagnetic field. The transition from examining symmetries of particles to the field is a big shift. We can't just consider fields as just functions of time, but as functions of both space and time. To do that, we need to introduce the mathematical concept of tensors.

To better understand the difference between Lagrangian mechanics in classical vs field theory, is to think of an analogy of a car traveling along a road. The Lagrangian allows us to describe the car's motion by incorporating its speed and position on the road. This is how we treat particles. But with fields, we instead think of multiple cars along the road, and each segment of the road can have unique characteristics and dynamics, and this is where the idea of Lagrangian density becomes important because they are functions of both space and time.

Think of each segment of the road as representing a point in space and time.

(Φ) could represent the number of cars or their speed at each point on the road. Aka. Traffic flow.

($\partial_\mu \Phi$) represents how the number of cars or their speed changes from one point on the road to another. Aka. Changes in the traffic.

So, the Lagrangian density describes the traffic dynamics over the entire road. It shows how each segment of the road interacts with its neighbors in space and time.

Another shift we will be considering, finding the true usefulness of Noether's theorem, is to instead examine how the existence of conservation laws leads to conserved quantities. We will be analyzing the conservation of electric charge and how it gives rise to a symmetry, specifically gauge symmetry.

Electromagnetic Field: The electromagnetic field is described by 4 equations, namely the Maxwell equations.

1. $\nabla \cdot E = \frac{\rho}{\varepsilon_0}$ -- Gauss's Law for Electricity
2. $\nabla \cdot B = 0$ -- Gauss's Law for Magnetism
3. $\nabla \times E = -\frac{\partial B}{dt}$ -- Faraday's Law
4. $\nabla \times B = \mu_0 J + \mu_0 \varepsilon_0 \frac{\partial E}{\partial t}$ -- Ampère's Law

Another way to describe this is to use the Electromagnetic Tensor. Which gives us the benefit of being invariant of the coordinate system we use. The Electromagnetic Tensor is defined as:

$$F_{\mu\upsilon} = \partial_\mu A_\upsilon - \partial_\upsilon A_\mu \tag{3.1}$$

Some understanding of the notation here is necessary before proceeding.

- A_μ represents the components of the electromagnetic potential,
- ∂_μ denotes the partial derivative with respect to the spacetime coordinate, x^μ
- μ, υ are indices running from 0 to 3, corresponding to the spacetime dimensions (time and spatial dimensions).

Conservation of charge: In any closed system, the sum of all positive and negative charges remains unchanged.

Consider Figure 6. The total charge inside is found by integrating the charge density over the volume:

$$Q(t) = \int_V \rho(r,t)\ d^3r \tag{3.2}$$

The change in Q over time is due to the flow of charge across the boundary surface B of V. Now, we introduce another variable, current density (J). This is a vector representing the amount of charge per unit area per unit time flowing across S.

Consider a patch on B at point r with are *dA* The current through this patch is given by the component of J perpendicular to the surface, where n is the unit vector normal to the surface: *J . n*

Current through the patch = *J . ndA*

Integrating this over the entire surface, we get,

$$I = \int_B J \cdot n\, dA \tag{3.3}$$

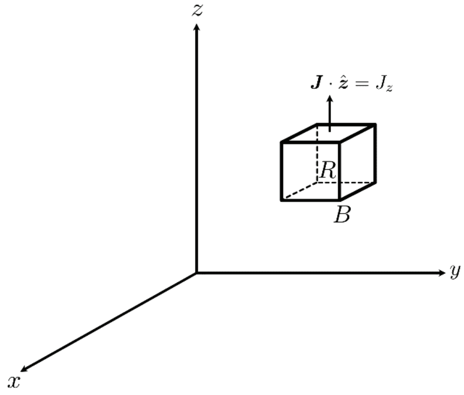


Figure 5: Volume *V* in space with charge density representing the charge per unit volume at point r and time t.

Now we can state the conservation of charge mathematically. (I) measures the amount of charge per unit time leaving the box (or entering it, if [I] came out negative). Local conservation of charge is the statement that if charge (I) per unit time flows out through the boundary, then the amount of charge Q inside the volume of the box goes down at that same rate:

$$\frac{dQ}{dt} = -I \tag{3.4}$$

The minus sign reflects our convention that I > 0 means outward flow. To convert the surface integral in (3.3) into a volume integral, we need to apply Gauss's theorem.

$$\oint_B J \cdot n\, dA = \int_V \nabla \cdot J\, d^3r \tag{3.5}$$

Substituting (3.5) into (3.4) and using $Q(t) = \int_V \rho\, d^3r$, we get:

$$\frac{d}{dt}\int_R \rho\ d^3r = -\int_B J \cdot n\, dA \tag{3.5}$$

To encompass *all* of space, so that the boundary is going to infinity, the current density (J) should go to zero in any physically reasonable setup, since there's nowhere left for the current to flow out to. Then the right-hand side vanishes, and this equation says that the total charge in all of space is constant.

Since this equation holds for any arbitrary volume (V), the integrands must be equal pointwise, leading to the continuity equation:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot J = 0 = \partial_0 \rho + \nabla \cdot J \tag{3.6}$$

The continuity equation in four-dimensional spacetime is:

$$\partial_\mu j^\mu = 0 \tag{3.7}$$

This depicts conservation of charge. Now we will go back to the electromagnetic field tensor, which is a key concept in relativistic electromagnetism. It compactly encapsulates the electric and magnetic fields.

$$F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu \tag{3.8}$$

Where A^μ is the four-potential, which includes the scalar Φ potential and the vector potential A

$$A^\mu = (\phi, A) \tag{3.9}$$

Now, to derive conservation of charge from the symmetry, we need the Lagrangian.

The Lagrangian density for the free electromagnetic field is given by:

$$L_{EM} = \frac{-1}{4} F_{\mu\nu} F^{\mu\nu} \tag{3.10}$$

The coupling of the electromagnetic field to charged particles is introduced through the current density J^μ and the four-potential A_μ. The coupling term in the Lagrangian density is:

$$L_{int} = -j^\mu A_\mu \tag{3.11}$$

The total Lagrangian density L is the sum of the electromagnetic field Lagrangian and the interaction term.

$$L = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} - j^\mu A_\mu \tag{3.12}$$

Now we perform a symmetrical operation that leaves the Lagrangian invariant. Such an operation or transformation is called the Gauge transformation.

$$A'_\mu = A_\mu + \partial_\mu \alpha \tag{3.13}$$

α is some spacetime scalar function. Let's now see how the Electromagnetic Tensor changes.

$$F'_{\mu\nu} = \partial_\mu A'_\nu - \partial_\nu A'_\mu \tag{3.14}$$

Substituting the transformed potential $A_\mu{}'$:

$$F'_{\mu\nu} = \partial_\mu (A_\nu + \partial_\nu \alpha) - \partial_\nu (A_\mu + \partial_\mu \alpha) = \partial_\mu A_\nu + \partial_\mu \partial_\nu \alpha - \partial_\nu A_\mu - \partial_\nu \partial_\mu \alpha \tag{3.15}$$

Since the mixed partial derivatives are symmetric,

$$\partial_\mu \partial_\nu \alpha = \partial_\nu \partial_\mu \alpha \tag{3.16}$$

These terms cancel out, leaving us with:

$$F'_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu = F_{\mu\nu} \tag{3.17}$$

So, the field tensor is invariant under the gauge transformation. So, the Lagrangian density should also be invariant. $L = L'$.

Noether's Procedure:

Now let's perform Noether's procedure with the gauge transformation. Since the Electromagnetic Tensor was invariant under Gauge transformation, the field term $L_{EM} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} = 0$

So, its variation in the action does not change. Let's look at the coupling terms. $L = -j^\mu A_\mu$

Applying a variation, we get: $\delta L = -j^\mu \partial A_\mu = -j^\mu A_\mu - -j^\mu \partial_\mu \alpha$.

$$\delta S = \int_{t_1}^{t_2} \delta L \; d^4x = - \int_{t_1}^{t_2} \left[j^\mu A_\mu + j^\mu \partial_\mu \alpha \right] d^4x \tag{3.18}$$

$\int_{t_1}^{t_2} j^\mu A_\mu$ is just the original action of the Lagrangian, which is $\Rightarrow - \int_{t_1}^{t_2} j^\mu A_\mu \, d^4x$ We now perform Integration by parts.

$$\Rightarrow - \int_{t_1}^{t_2} j^\mu A_\mu \, d^4x = \left[j^\mu A_\mu \right]_{t_1}^{t_2} - \int_{t_1}^{t_2} j^\mu \partial_\mu \alpha \, d^4x \tag{3.19}$$

$\alpha(t_1) = \alpha(t_2) = 0$. From the conservation of charge, we have:

$$\partial \mu J^u = 0 \Rightarrow \int_{t_1}^{t_2} \alpha(\partial \mu J^\mu) \, d^4x \tag{3.20}$$

Because of the current conservation ($\partial_\mu J^\mu$), this vanishes and $\delta S = 0$. Showing the Gauge symmetry. But there are some important implications. We assumed that conservation of charge had to exist for there to be Gauge symmetry. We didn't get any conserved quantity. This is because we had a redundancy of information. This means that all components of a field tensor or set of equations are not independent. Some can be derived from others due to symmetries or constraints. This redundancy ensures that physical principles, like the conservation of electric charge, are naturally satisfied and can't be derived directly from Noether's theorem.

4: Noether's theorem to QFT:

Symmetries and Noether's theorem can also be applied to quantum fields. In quantum field theory, every type of particle is associated with a corresponding quantum field. For example, the electromagnetic field is associated with photons, while the electron field is associated with electrons. The specific field we will be looking at is the complex scalar field.

4.1: Complex scalar field:

Unlike the electromagnetic field tensor, which describes both the electric and magnetic fields, a complex scalar field is a type of quantum field characterized by values that are complex numbers (numbers that have both real and imaginary parts) at each point in space and time. An example of a complex scalar field is the Higgs Field. The Lagrangian density of a complex scalar field:

$$L = (\partial_\mu \phi^*)(\partial_\mu \phi) - m^2 \phi^* \phi \tag{4.1}$$

DOI: 10.36838/v8i2.39

We now perform a global U (1) phase transformation.
$\Phi \rightarrow \Phi' = e^{iß}\,\Phi$, $\Phi^* \rightarrow \Phi^{*'} = e^{iß}\,\Phi^{*'}$

This changes the phase of the field across all spacetime points. In the transformation *ß*, is a constant function and $e^{iß}$ is a complex number that changes the phase of *Φ(x)* and *Φ*(x)*. To see the Lagrangian is symmetric under this transformation, first let's look at the kinetic energy term:

$$\partial_\mu \phi \;\rightarrow\; \partial_\mu \phi' = \partial_\mu\left(e^{i\beta}\,\phi\right) = e^{i\beta}\,\partial_\mu \phi \qquad (4.2)$$

$$\partial_\mu \phi^* \;\rightarrow\; \partial_\mu \phi^{*'} = \partial_\mu\left(e^{-i\beta}\,\phi^*\right) = e^{-i\beta}\,\partial_\mu \phi^* \qquad (4.3)$$

The product of these two terms becomes:

$$\left(\partial_\mu \phi^*\right)\left(\partial_\mu \phi\right) \;\rightarrow\; \left(e^{-i\beta}\,\partial_\mu \phi^*\right)\left(e^{i\beta}\,\partial_\mu \phi\right) = 1 \cdot \partial_\mu \phi^* \partial_\mu \phi \qquad (4.4)$$

This is because the phase factors $e^{-iß}$ and $e^{iß}$ α cancel each other out, leaving the kinetic term unchanged. To see why this is, we'll analyze Euler's formula.

$$e^{i\beta} = \cos(\beta) \;+ i\sin(\beta) \text{ and } e^{-i\beta} = \cos(-\beta) \;+ i\sin(-\beta) = \cos(\beta) \;- i\sin(\beta) \qquad (4.5)$$

Multiplying them together, we get.

$$e^{i\beta} \cdot e^{-i\beta} = [\cos(\beta) \;+ i\sin(\beta)][\cos(\beta) \;- i\sin(\beta)] = \cos^2(\beta) - i\cos(\beta)\sin(\beta) + i\sin(\beta)\cos(\beta) - (i\sin(\beta))^2 = \cos^2(\beta) - (i\sin(\beta))^2$$

and since

$$i^2 = -1, \Longrightarrow \;\; = \cos^2(\beta) + \sin^2(\beta) \;\; = 1 \qquad (4.6)$$

Now let's analyze the mass term.

$$\phi^*\phi \rightarrow \phi'^*\phi' = \left(e^{-i\beta}\phi^*\right)\left(e^{i\beta}\phi\right) = \phi^*\phi \qquad (4.7)$$

Since the 2 phases cancel out, we get back the original *Φ* terms. The mass will stay constant, and therefore the mass term is also invariant, meaning the whole Lagrangian is invariant. *L =L'.*

<u>Noether's procedure:</u>

$$\delta[S] = \int_{t_1}^{t_2} \delta L \;\, d^4x \qquad (4.8)$$

Where $\delta L = \frac{\partial L}{\partial \phi}\delta\phi + \frac{\partial L}{\partial \phi^*}\phi^* + \partial\frac{\partial L}{(\partial_\mu \phi)}\delta\left(\partial_\mu \phi\right) + \frac{\partial L}{\partial(\partial_\mu \phi^*)}\delta\left(\partial_\mu \phi\, *\right)$ (4.9)

Now let's solve each term by term of the variation in the Lagrangian.

$$\frac{\partial L}{\partial(\partial_\mu \phi)}\delta\left(\partial_\mu \phi\right) = \partial^\mu \phi^* \quad \text{and} \quad \frac{\partial L}{\partial(\partial_\mu \phi^*)}\delta\left(\partial_\mu \phi^*\right) = \partial^\mu \phi^* \qquad (4.10)$$

$\delta\phi \;=\; \phi' - \phi \;=\; e^{i\beta}\phi \;-\; \phi$ We can expand $e^{iß}$ using Taylor's expansion.

$$e^{i\beta} = 1 + i\beta - \frac{\beta^2}{2!} + \cdots \;\Rightarrow\; e^{i\beta}(\phi) = \phi + i\beta\phi - \frac{\beta^2}{2!}\phi + \cdots \;\Rightarrow\; e^{i\beta}(\phi) - \phi = i\beta\phi - \frac{\beta^2}{2!}\phi + \cdots$$

We can neglect the higher-order terms and be left with $\delta\phi \;\approx\; i\beta\phi$ and $\delta\phi^* \;\approx\; -\,i\beta\phi^*$

Also notice that $\delta(\partial_\mu \phi) = \partial_\mu(\delta\phi) = \partial_\mu(i\beta\phi)$ and $\delta\left(\partial_\mu \phi^*\right) = \partial_\mu(\delta\phi^*) = \partial_\mu(-i\beta\phi^*)$

Lastly, we have, $\frac{\partial L}{\partial \phi} = -m^2\phi^*$ and $\frac{\partial L}{\partial \phi^*} = -m^2\phi^*$. Now let's substitute these values into the variational Lagrangian.

$$\delta L = -m^2\phi^*\phi i\beta + m^2\phi\phi^* i\beta + \;\partial^\mu\phi^*\partial_\mu\phi i\beta - \partial^\mu\phi\partial_\mu\phi^* i\beta \;=\; i\beta\left(\phi^*\partial^\mu\partial_\mu\phi - \phi\partial^\mu\partial_\mu\phi^*\right)$$

$$\delta[S] = \int_{t_1}^{t_2} \delta L \;\, d^4x = \int_{t_1}^{t_2}\left(\, i\beta\phi^*\partial^\mu\partial_\mu\phi\right)d^4x - \int_{t_1}^{t_2}\left(i\beta\phi\partial^\mu\partial_\mu\phi^*\right)d^4x \qquad (4.11)$$

We can now perform integration by parts on the first integral and plug in the endpoints.

$\int_{t_1}^{t_2}\left(\, i\beta\phi^*\partial^\mu\partial_\mu\phi\right)d^4x \;=\; i\left(\left[\beta\partial_\mu\right]_{t_1}^{t_2} - \int_{t_1}^{t_2}\beta\phi^*\,\partial^\mu\phi\right)$. Since $\beta(t_1) = \beta(t_2) = 0$, we are left with:

$$-i\int_{t_1}^{t_2}\beta\phi^*\,\partial^\mu\phi \qquad (4.12)$$

We can do the same for the second integral. Perform integration by parts and plug in endpoints.

$\int_{t_1}^{t_2}\left(\, i\beta\phi^*\partial^\mu\partial_\mu\phi\right)d^4x \;=\; -i\left(\left[\beta\partial_\mu\right]_{t_1}^{t_2} - \int_{t_1}^{t_2}\beta\phi\partial^\mu\,\phi^*\right)$. since $\beta(t_1) = \beta(t_2) = 0$, we are left with:

$$i\int_{t_1}^{t_2}\beta\phi\partial^\mu\,\phi^* \qquad (4.13)$$

Combining these two integrals together, we have,

$$i\int_{t_1}^{t_2}\beta[\partial^\mu\phi\phi^* - \partial^\mu\phi^*\phi]d^4x \;=0 \qquad (4.14)$$

By the principle of least action

$$\Longrightarrow i[\partial^\mu\phi\phi^* - \partial^\mu\phi^*\phi] \;=0 \qquad (4.15)$$

So, the conserved current is J^μ = i $[\partial^\mu\Phi\Phi^* - \partial^\mu\Phi^*\Phi]$. This depicts the conservation of charge in QED.

What if *ß* was instead a function of space and time. *ß* → *α (x, t)*

$\Phi \rightarrow \Phi' = e^{i\alpha(x,t)}\Phi$ and $\Phi^* \rightarrow \Phi^{*'} = e^{-i\alpha(x,t)}\Phi^*$. Like before, let's analyze the K.E. term first.

$\partial_\mu\phi \rightarrow \partial_\mu\left(e^{i\alpha(x,t)}\phi\right) = e^{i\alpha(x,t)}\left(\partial_\mu\phi + i(\partial_\mu\alpha)\phi\right)$. We get this expression if we use Taylor's Expansion. We could do the same for $\partial_\mu\Phi^*$.

$$\partial_\mu\phi^* \;\rightarrow \partial_\mu\left(e^{i\alpha(x,t)}\phi^*\right) = e^{i\alpha(x,t)}\left(\partial_\mu\phi^* - i\left(\partial_\mu\alpha\right)\phi^*\right). \qquad (4.16)$$ [14]

Therefore, $L' = e^{i\alpha(x,t)} \cdot e^{-i\alpha(x,t)}\left(\partial_\mu\phi^* - i\left(\partial_\mu\alpha\right)\phi^*\right)\left(\partial_\mu\phi + i\left(\partial_\mu\alpha\right)\phi\right) - m^2\phi'\phi'^*$. Expanding, we get.

$$L' = \partial_\mu\phi^*\partial^{\{(\mu}\phi \;+\; i\phi^*\left(\partial_\mu\alpha\right)\partial^{\{(\mu}\phi - i\phi(\partial^\mu\alpha)\partial_\mu\phi^* - \left(\partial_\mu\alpha\right)(\partial^\mu\alpha) + (\phi^*\phi) - m^2\phi'\phi'^* \qquad (4.17)$$

This shows that *L* ≠ *L′* because of the additional: $\partial_\mu\alpha$ terms. They don't vanish.

When the phase transformation was a function of space and time, the Lagrangian of the complex scalar wasn't invariant. This is called the local U (1) phase transformation. This transformation led to a symmetry that was internal. The phase of the complex scalar field can be locally altered (meaning it can be changed at each point independently) without affecting the overall physics. The specific conserved quantity of a complex scalar field depends on the context of the field and the physical theory being discussed. In some theories, complex scalar fields can represent particles with conserved quantum numbers like baryon number or lepton number. In other theoretical models, such as quantum electrodynamics (QED), a complex scalar field can represent particles with electric charge. It can also be applied to the Standard Model of particle physics, where the Higgs field is a complex scalar field.[15]

This invariance under the local phase changes is important in constructing theories like the Standard Model of particle physics. This also leads to the Higgs field and spontaneous symmetry breaking. The Higgs field has a symmetric potential, and the field value at each point is zero, which represents a high-energy, unstable state. When the field transitions to a lower-energy state where it has a non-zero magnitude everywhere in space, the specific direction of the field in the complex plane breaks the symmetry. Particles interacting with the Higgs field will then gain mass. The interaction depends on the field's magnitude, which is now non-zero and uniform across space.[16]

We will look at the complex scalar field in the context of QED.

4.2: Scalar QED:

Let's now apply Noether's theorem to scalar QED, which extends the principles of QED as it deals with spin-1/2 particles like electrons to scalar fields. Scalar QED is an extension of classical electrodynamics and quantum field theory. It describes the interaction between scalar fields (fields that are represented by scalar particles, which have spin zero) and the electromagnetic field.[17] The Lagrangian of the scalar QED is:

$$L = \left(D_\mu\phi\right)^*\left(D_\mu\phi\right) - m^2\phi^*\phi\ - \tfrac{1}{4}F_{\mu\nu}F^{\mu\nu} \qquad (4.18)$$

The mass term and the electromagnetic tensor have already been introduced before. The only new thing is the covariant derivative, incorporating the interaction with the gauge field A_μ.[18]

$$L = \left(\partial_\mu\phi^* + ieA_\mu\phi\,*\right)\left(\partial^u\phi - ieA^\mu\phi\right) - m^2\phi * \phi - \tfrac{1}{4}F_{\mu\nu}F^{\mu\nu}\ (4.19)$$

Let's apply the local U (1) phase transformation:

$$\phi^* \to \phi^{*'} = e^{-i\alpha(x,t)}\phi^*$$

$D_\mu\phi \to D'_\mu\phi'(\partial_\mu - ieA_\mu)e^{-i\alpha(x,t)}\phi.$ We can then use the product rule to obtain:

$$D_\mu\phi = (\partial_\mu - ieA_\mu)e^{-i\alpha(x,t)}\phi = e^{i\alpha(x,t)}\left(\partial_\mu + i(\partial_\mu\alpha)\right)\phi - ieA_\mu e^{i\alpha}\phi = e^{i\alpha(x,t)}(\partial_\mu\phi + i(\partial_\mu\alpha)\phi) - ieA_\mu\phi =$$
$$e^{i\alpha(x,t)}\left(\partial_\mu - ie\left(A_\mu - \tfrac{1}{e}\partial_\mu\alpha\right)\right)\phi.$$

To make sure the covariant derivative is invariant, the gauge field A_μ must transform in the way obtained in the double parentheses. $A_\mu \to A'_\mu\ = A_\mu + \frac{1}{e}\partial_\mu\alpha$

This makes sure that $D_\mu\phi\ \to\ D'_\mu\phi' = e^{i\alpha}D_\mu\phi.$ So now we can combine the two terms and get:

$$(D_\mu\phi^*)\,(D_\mu\phi\,) \to (D'_\mu\phi'^*)\,(D'_\mu\phi'\,) = e^{i\alpha}(D_\mu\phi^*)e^{-i\alpha}(D_\mu\phi) = (D_\mu\phi^*)(D_\mu\phi). \text{ Since } e^{i\alpha}\cdot e^{-i\alpha} = 1 \qquad (4.20)$$

The same applies to the mass term.

$$m^2\phi^*\phi \to m^2(e^{-i\alpha}\phi^*)(e^{i\alpha}\phi) = m^2\phi^*\phi$$

What about $F_{\mu\nu}$? $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. Remember the transformation we obtained to make the covariant derivative invariant:

$$A_\mu \to A'_\mu = A_\mu\ + \tfrac{1}{e}\partial_\mu\alpha \Longrightarrow F'_{\mu\nu} = \partial_\mu A'_\nu - \partial_\nu A'_\mu = \partial_\mu\left(A_\nu\ + \tfrac{1}{e}\partial_\mu\alpha\right) - \partial_\nu\left(A_\mu\ + \tfrac{1}{e}\partial_\mu\alpha\right). \qquad (4.21)$$

Since $\partial_\mu\partial_\nu\alpha = \partial_\nu\partial_\mu\alpha \Longrightarrow F'_{\mu\nu} = F_{\mu\nu}.$ and every term of the scalar QED is invariant, the whole scalar QED is invariant under the local U (1) phase transformation. Let's now get Noether's current for this. The transformations are

$$\phi\ \to\ \phi' = e^{i\alpha(x,t)}\phi,\ \ \phi^* \to \phi^{*'} = e^{-i\alpha(x,t)}\phi^*,\ A_\mu \to A'_\mu = A_\mu - \tfrac{1}{e}e\partial_\mu\alpha \qquad (4.22)$$

The variation in the Lagrangian is:

$$\delta L = \frac{\partial L}{\partial\phi}\delta\phi + \frac{\partial L}{\partial\phi^*}\delta\phi^* + \frac{\partial L}{\partial(\partial_\mu\phi)}\delta(\partial_\mu\phi) + \frac{\partial L}{\partial(\partial_\mu\phi^*)}\ \delta(\partial_\mu\phi^*) + \frac{\partial L}{\partial A_\nu}\delta A_\nu + \frac{\partial L}{\partial(\partial_\mu A_\nu)}\ \delta(\partial_\mu A_\nu)(4.23)$$

Integrating the derivative-variation terms by parts at the density level and collecting total derivatives, this separates the pieces that are proportional to the equations of motion from a total divergence:

$$\delta L = \left(\frac{\partial L}{\partial\phi} - \partial_\mu\frac{\partial L}{\partial(\partial_\mu\phi)}\right)\delta\phi + \left(\frac{\partial L}{\partial\phi^*} - \partial_\mu\frac{\partial L}{\partial(\partial_\mu\phi^*)}\ \right)\delta\phi^* + \left(\frac{\partial L}{\partial A_\nu} + \partial_\mu F^{\mu\nu}\right)\delta A_\nu$$
$$+\ \partial_\mu\left[\frac{\partial L}{\partial(\partial_\mu\phi)}\delta\phi + \frac{\partial L}{\partial(\partial_\mu\phi^*)}\delta\phi^* - F^{\mu\nu}\delta A_\nu\right] \qquad (4.24)$$

Now to specialize the variations to the infinitesimal local U(1) phase:

$$\delta\phi = i\alpha(x)\phi,\ \delta\phi^* = -\,i\alpha(x)\phi^*,\ \ \delta A_\nu = \frac{1}{e}\partial_\nu\alpha(x). \qquad (4.25)$$

(For small α we use $e^{\pm i\alpha} \approx i\alpha$.)

Because the Lagrangian is gauge invariant, $\delta S = 0$ for these variations. Next, we take the global subset of the symmetry by setting α constant, so $\delta A_v = 0$. Substituting (4.24) into (4.23) and using $\delta A_v = 0$. gives:

$$0 = \left(\frac{\partial L}{\partial\phi} - \partial_\mu\frac{\partial L}{\partial(\partial_\mu\phi)}\right)i\alpha\phi + \left(\frac{\partial L}{\partial\phi^*} - \partial_\mu\frac{\partial L}{\partial(\partial_\mu\phi^*)}\right)(-i\alpha\phi^*) +\ \partial_\mu\left[i\alpha\left(\frac{\partial L}{\partial(\partial_\mu\phi)}\phi - \frac{\partial L}{\partial(\partial_\mu\phi^*)}\phi^*\right)\right] \qquad (4.26)$$

The two parenthetical factors multiplying $i\alpha\Phi$ and ($-i\alpha\Phi^*$) are exactly the left-hand sides of the Euler-Lagrange equations for Φ and Φ^*. When the fields satisfy their equations of motion, those factors will vanish. Dropping that term and removing the overall constant α leads to the local continuity equation:

$$\partial_\mu\left(i\phi\frac{\partial L}{\partial(\partial_\mu\phi)} - \ i\phi^*\frac{\partial L}{\partial(\partial_\mu\phi^*)}\right) = 0 \qquad (4.27)$$

Therefore, the conserved Noether current is:

$$J^\mu = i\phi\frac{\partial L}{\partial(\partial_\mu\phi)} - \ i\phi^*\frac{\partial L}{\partial(\partial_\mu\phi^*)}, \qquad \partial_\mu J^\mu = 0. \qquad (4.28)$$

Finally, using the known expressions for the derivative terms in scalar QED,

$$\frac{\partial L}{\partial(\partial_\mu\phi)} = (D^\mu\phi)^*, \qquad \frac{\partial L}{\partial(\partial_\mu\phi^*)} = (D^\mu\phi) \qquad (4.29)$$

To write the current in the familiar form,

$$J^\mu = i(\phi(D^\mu\phi)^* - \phi^* D^\mu\phi), \qquad \partial_\mu J^\mu = 0 \qquad (4.30)$$

■ Discussion

Noether's theorem is one of the most useful tools for theoretical physics. It has applications wherever there is continuous symmetry in any physical system. Whether the system is local or isolated, the conservation laws that are derived from the symmetry and valid exactly and can be easily applied to simplify problems in both classical and quantum physics.

The importance of Noether's theorem extends even further when we consider different frames of reference. The laws of physics must be invariant in different frames. This requires the introduction of extra structure to maintain that invariance. For example, in non-inertial frames, fictitious forces like centrifugal or Coriolis forces are introduced to make sure that Newton's

DOI: 10.36838/v8i2.39

laws are valid in these systems. Similarly, in particle physics, gauge fields are introduced to maintain invariance under local symmetries. Specifically, invariance under local phase shifts in the quantum field of the electron involves introducing the electromagnetic field, which naturally couples to the electric charge.

This principle of extra fields arising to maintain local symmetries gives us valuable insight into reality. They hint at the existence of fundamental interactions like electromagnetism, and the reason there is conservation of electric charge and the existence of light. This principle is also at the bedrock of particle physics. Quarks within protons and neutrons follow a symmetry based on the number three, while discrete symmetries such as charge conjugation (C), parity (P), and time reversal (T) give us valuable insight into understanding particles and anti-particles.

We must, however, be careful not to overextend the theorem into areas it can't. To point to its biggest limitation, the theorem breaks down when there are only discrete symmetries or no symmetry at all. An example of this is where spacetime itself is dynamical. In that case, the underlying symmetries don't hold globally. We see this with the expansion of the universe, which breaks perfect time-translation symmetry. As a result, energy isn't conserved at the cosmic level. This is evidence when we look at the redshift of light, where photons lose energy as their wavelengths stretch with the expanding universe. Similarly, the universe also doesn't have perfect spatial symmetry because of the unequal distribution of stars, planets, and other structures. This implies that we can't apply conservation of momentum globally.

Research around this subject is constantly being done to find out if there are more fundamental symmetries. One popular domain of research is Supersymmetry, which suggests that there might be a deep symmetry between matter particles and force-carrying particles, pointing to a unified framework for the forces of our universe. Whether or not these symmetries hold in nature is still under research, but symmetries and Noether's theorem are at the forefront of shaping modern physics.

■ Conclusion

As we have shown, Noether's theorem is applicable across many systems in physics. Even in cases such as the gauge symmetry, where there is a redundancy of information, the conserved quantity had to be necessarily true for the Lagrangian to be invariant.

There were many limitations to this paper. As mentioned in the methods, there was a limitation in deriving the Lagrangian of the complex systems. Moreover, many specific cases and extensions of Noether's theorem have not been considered. Noether's theorem assumes that any symmetry under consideration must be continuous. Even though no conserved quantity is derived from such discrete symmetries, they often impose selection rules in quantum systems, which limit possible transitions or interactions. There are many examples, such as Parity Symmetry, Time Reversal Symmetry, and Charge Conjugation Symmetry.

While Noether's theorem and its applications have been well established, there is still much ongoing research concerning its implications and reach. As discussed, there is still ongoing research about spontaneous symmetry breaking. Research continues into how symmetries can be spontaneously broken in various physical systems, which leads to phenomena such as the Higgs mechanism. There can also be extensions in the formalism of discrete symmetries, which do not lead to conserved quantities but can have many physical implications. Of course, there are important considerations that need to be made, such as to what extent Noether's theorem can be applied outside of physics, but that is a question for further research.

■ Acknowledgments

I would like to give my deepest gratitude to my professor and mentor, Dr. Surjeet Rajendran of Johns Hopkins University, who guided, evaluated, and supported me throughout this research project. I am very thankful to my PCA, Ramzan Ullah, for the help and support provided during the Pioneer Academics program. I would also like to acknowledge the contribution of my peer reviewer, Enzo Visintin, whose review and feedback greatly improved the quality of this paper. I also stand highly indebted to Nowyshin Mridula and her friend Kevin Tanious, both from Yale University, who put in diligent efforts to edit and proofread my paper. Their careful attention to detail was greatly appreciated. My thanks also extend to my Ligaya Cullo for their support in editing and giving feedback. Finally, I would like to thank Pioneer Academics for this opportunity to give resources and a platform to conduct research.

■ References

1. Robinson, M. *Symmetry and the Standard Model: Mathematics and Particle Physics*; Springer: New York, 2011.
2. Maudlin, T., Okon, E., Callender, C., Pérez, D. *On the Status of Conservation Laws in Physics: Implications for Semiclassical Gravity. Stud. Hist. Philos. Sci. B* 2020, 69, 67–81. DOI: 10.1016/j.shpsb.2019.10.004.
3. Newton, I.; Motte, A.; Cajori, F. *Mathematical Principles of Natural Philosophy and His System of the World*, Vol. 1; University of California Press: Berkeley, 1966.
4. Goldstein, H. *Classical Mechanics*, 2nd ed.; Wesley: Reading, MA, 1980.
5. Strogatz, S. H. *Nonlinear Dynamics and Chaos*, 2nd ed.; Westview Press: Boulder, CO, 2015.
6. Landau, L. D.; Lifshitz, E. M. *Mechanics*, 3rd ed.; Elsevier: Oxford, 1982.
7. Lanczos, C. Rev. Mod. Phys. **1949**, 21, 497–502.
8. Gelfand, I. M.; Fomin, S. V. *Calculus of Variations*; Courier Corporation: New York, 2012.
9. Lagrange, J. L.; Todhunter, I.; Whittaker, E. *Analytical Mechanics*; Kluwer Academic Publishers: Dordrecht, 1997.
10. Noether, E. Invariant Variation Problems. *Transp. Theory Stat. Phys.* **1971**, 1(3), 186–207. DOI: 10.1080/00411457108231446.
11. Kosmann-Schwarzbach, Y. *The Noether Theorems*; Springer: Berlin, 2010.
12. Ryder, L. H. *Quantum Field Theory*, 2nd ed.; Cambridge University Press: Cambridge, 1985.
13. Krane, K. S. *Modern Physics*; John Wiley & Sons: Hoboken, NJ, 2020.

14. Peskin, M. E.; Schroeder, D. V. *An Introduction To Quantum Field Theory*; Westview Press: Boulder, CO, 1995.
15. Griffiths, D. J. *Introduction to Elementary Particles*, 2nd ed.; Wiley-VCH: Weinheim, Germany, 2008. (For Higgs field and Standard Model).
16. Greiner, W.; Reinhardt, J. *Field Quantization*; Springer: Berlin, Heidelberg, 1996. (For spontaneous symmetry breaking and Higgs mechanism).
17. Srednicki, M. *Quantum Field Theory*; Cambridge University Press: Cambridge, 2007. (General QFT reference, including scalar QED).
18. Mandl, F.; Shaw, G. *Quantum Field Theory*, 2nd ed.; John Wiley & Sons: Chichester, England, 2010. (Specifically for covariant derivative and gauge interactions).

■ Author

Hasin A. Shaykat is a junior at The Kew Forest School. He plans to pursue a major in Physics or Applied Mathematics in college, with a keen interest in theoretical physics and its foundational principles.

DOI: 10.36838/v8i2.39

IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Cascade Use or Recycle? Secondary Life Planning for Electric Vehicle Batteries in China

Jeffrey J. Shen

West Island School, 250 Victoria Road, Hong Kong Island, Hong Kong; jeffreyshen2009@gmail.com

ABSTRACT: China's electric vehicle (EV) industry has experienced rapid growth in recent years, becoming a significant driver of economic and technological advancements. Lifecycle management of EV batteries is now a pressing issue due to environmental concerns. This paper investigates this problem, focusing on the dual strategy of recycling and cascade utilization. Using extensive real-world data, models are estimated to predict EV batteries' performance and lifespan under practical conditions. These capacity degradation models are then applied to forecast the future growth of EV and non-electric vehicle (non-EV) battery volumes and the capacity structure of the battery population. Furthermore, a mathematical model is formulated to describe the flow of batteries in the EV and non-EV population, capturing EV battery transitions through cascade utilization to non-EV uses and to recycling. An optimization problem is then proposed to maximize social utility and to guide decisions on cascade utilization.

KEYWORDS: Environmental Engineering, Recycling and Waste Management, Electric Vehicle Batteries, Management Science.

Introduction

Over the past five years, Electric Vehicles (EVs) have seen rapid growth globally. In 2024, EV sales exceeded 17 million, accounting for over 20% of all vehicle sales. The European Union, the United States, and China remain the three major markets, and Asia and Latin America are important emerging markets that see rapid growth in EV sales.[1]

The forecast for 2025 shows that EV sales are expected to grow further to exceed 20 million worldwide. The driving force for such growth varies in different markets. For example, in China, it is policy incentives, such as a trade-in scheme where higher rebate is offered for the purchase of an EV purchase than that the purchase of a conventional vehicle, as well as infrastructure development, and domestic manufacturing capacity, that will push share of EV sales up to 60%; in Europe, it is the emission reduction target that will drive up the shares of zero-emission EVs to 25%; and in the United States, sales are projected to raise slightly to 11% due to change in policy direction.[2]

Driven mostly by the increase in EV sales, in 2024, battery production that satisfies both EV demand and storage applications reached the 1TWh milestone. China remains the largest source of demand at 60% of global demand, European Union and the United States at 13%.[2]

Recognizing these international differences, this paper focuses on the Chinese context, examining optimal recycling and cascade utilization strategies tailored to its unique market conditions and policy environment.

Beyond economic benefits, the widespread adoption of EVs also brings substantial environmental advantages. Electrification of transportation is seen as a pivotal strategy to reduce dependence on petroleum-based fuels and to mitigate urban air pollution.[3] EVs, with zero tailpipe emissions, contribute to improved air quality and reduced greenhouse gas emissions when powered by low-carbon energy sources.

However, the rapid expansion of the EV market also brings new challenges, particularly in the management of retired batteries. With the rapid development of EVs, the number of retired batteries is expected to surge in the coming years. According to Wu *et al.*,[4] the volume of retired power batteries is projected to rise from 112,000 tonnes in 2020 to 708,000 tonnes by 2030. The substantial increase in retired batteries underscores the urgent need for efficient reuse and recycling strategies. Improper disposal of batteries can lead to severe environmental and safety risks. For instance, leaked heavy metals from improperly disposed batteries can contaminate water and soil, and pose threats to ecosystems and human health through bioaccumulation in the food chain.[5] Additionally, improper dismantling of EV batteries could pose significant safety concerns, such as fire or explosion.[6]

Typically, there are two main strategies for handling retired EV batteries: recycling and cascade utilization. Recycling involves dismantling batteries to reclaim valuable materials such as lithium, cobalt, and nickel, which can be used to manufacture new batteries. In other words, recycling aims to convert power batteries into various raw materials with minimal pollution. Cascade utilization, on the other hand, repurposes batteries for secondary applications after their initial use as EV batteries. This strategy extends the lifecycle of EV batteries and mitigates the environmental impact of battery disposal. Cascade utilization includes applications in energy storage systems,[7] backup for base stations,[8] grid support services,[9] and renewable energy integration,[10] etc. According to the report by the China Electricity Council,[11] from 2019 to 2022, storage demand grew from 466 MWh to 5,498 MWh for renewable energy stations, from 523 MWh to 1,812 MWh for the power network, and from 119 MWh to 758 MWh for commercial demand. The

DOI: 10.36838/v8i2.50

rapid growth demonstrates the potential need for cascade-utilized EV batteries.

While recycling is a straightforward solution, it fails to fully harness the capacity of EV batteries. From the perspective of social welfare, cascade utilization is a superior strategy as it enables maximal utilization of battery capacities. However, the proper planning, management, and operations of cascade utilization remain challenging.

In this paper, we analyze and explore issues related to the life cycle management of EV batteries, including: when to recycle and when to cascade utilize? How to balance the two strategies? How to develop them in the long term?

We first investigate the capacity degradation pattern for individual batteries to lay the foundation for our study. The modeling utilizes two publicly available datasets to provide insights into the typical lifespan and performance decline of EV batteries. The analysis is then extended from individual batteries to the entire battery population, depicting its capacity distribution at any point in time. A cascade utilization flow model is introduced to capture the transition of batteries from EV to non-Electric Vehicle (non-EV) markets, and their eventual flow to recyclers. Based on the current state of the EV and the cascade utilization market, we project future growth in battery volumes. It shows that the number of retired EV batteries will be enormous, with only a small fraction absorbed by the current cascade utilization market, thus highlighting a significant underutilization of this potential strategy.

In summary, this paper provides a quantitative analysis of the economic and policy factors influencing the cascade utilization of EV batteries. Through detailed modeling of battery degradation patterns, market projections, and the effects of government subsidies, this paper aims to inform and guide policymakers and industry stakeholders in making strategic decisions that enhance the sustainability and economic viability of EV battery lifecycle management.

■ Literature Review

The related literature mainly consists of research in two areas: the pattern of battery capacity degradation and multi-party relationships related to the recycling of batteries.

Capacity Degradation of Batteries:

Battery capacity degradation is a significant concern for the sustainability and performance of EVs. Different approaches have been used in modeling battery degradation. The first is by simulating the underlying physical degradation mechanisms. Edge *et al*. provide a comprehensive overview of lithium-ion battery degradation mechanisms.[12] They discuss the coupling between different degradation processes and propose a semi-empirical model that integrates physical and chemical degradation mechanisms. This model aims to predict capacity fade and enhance battery management systems. Luo *et al*. present a detailed study on capacity degradation and aging mechanisms in lithium-ion batteries under various operating conditions.[13] Their empirical model considers factors such as the solid electrolyte interphase (SEI) growth, lithium plating, and particle cracking to predict battery lifespan under different depths of discharge and temperatures.

Another approach is to employ data-driven methods to model the degradation process of batteries. Zhang *et al*. built an accurate battery forecasting system based on electrochemical impedance spectroscopy.[14] A Gaussian process model takes the entire collected spectrum as input and automatically determines which spectral features better predict degradation. Huang *et al*. propose a novel charging encoder that alternates between a Temporal Convolutional Network and a Bidirectional Gated Recurrent Unit to capture local temporal information and long-term dependencies related to the state of capacity (SOC) and the state of health (SOH) during charging.[15] The proposed framework enables a unified joint estimation of the two variables, substantially enhancing efficiency.

Recycling of EV Batteries:

As the battery recycling and cascade utilization market expands, more research efforts start to focus on the decision-making relationship between the various parties in this context.

Some of them focus on the strategy analysis of different roles in the supply chain, including pricing, contracts, and benefit distribution. Gu *et al*. propose a closed-loop supply chain model in which EV batteries can be reused, such as for energy storage, before being recycled.[16] They analyze the optimal pricing strategy between the manufacturer and remanufacturer to optimize the total profit in the whole supply chain. Zhu and Yu study the effect of adverse selection and moral hazards in the closed-loop supply chain of EV batteries based on Information Screening Models in the principal-agent theory.[17]

Some papers examine the impact of government policies. Gu *et al*. look for the optimal production strategy when market demand is uncertain under government subsidy.[18] It is concluded that the optimal production quantity and expected utility increase with the subsidy. Guan and Hou study the equilibrium strategy of the EV battery supply chain under the dual mechanism of government subsidy and cost-sharing and find that the utility of cascade utilization efforts will increase with the increase of government subsidies.[19]

In this paper, the focus is not on the benefits and decisions of participants at the micro level; instead, it focuses on the circulation of batteries from a macro perspective and hopes to optimize social welfare through macro-control measures such as cascade utilization standards.

■ Methods

Electric Vehicle Battery Capacity Degradation Model:

To develop effective recycling and cascade utilization strategies, it is essential to understand the mechanisms behind the capacity degradation of EV batteries over time. In this section, we propose an EV battery capacity degradation model to help accurately predict their lifespan and performance under real-world conditions.

Battery capacity refers to the total amount of electric charge a battery can store, quantified as a real number and measured in ampere-hours (Ah). Generally, a larger battery capacity al-

DOI: 10.36838/v8i2.50

lows for more energy storage, enabling EVs to travel greater distances on a single charge. It serves as a crucial indicator of a battery's health status, with higher values representing better overall performance.

The performance of EV batteries inevitably degrades with increased usage. This degradation is primarily reflected in the gradual decline of battery capacity. Over time, this reduction in capacity diminishes the battery's ability to store and deliver energy effectively. This degradation also forms the basis for cascade utilization. As EV batteries degrade over time, they eventually become unsuitable as power batteries but retain value for other applications. The timing of their retirement is critical to determining their remaining utility in secondary applications.

To characterize the overall condition of EV batteries, it is necessary to accurately describe the capacity degradation of EV batteries. Some studies have examined the performance of batteries under laboratory conditions. However, in real-world scenarios, the use of EV batteries is far more complex than under laboratory testing conditions. EV batteries are affected by numerous complex real-world factors, such as unstable voltage, random charging times, EV owners' charging preferences (charging only when nearly depleted or frequent partial charging), constantly changing ambient temperature, and more. These factors can render the battery degradation curves obtained under laboratory conditions invalid.

Therefore, to accurately model the capacity degradation of EV batteries, it is essential to build the model with extensive real-world data, which ultimately helps capture the most fundamental degradation pattern.

In our study, two publicly available EV battery datasets are utilized to model the degradation curve and perform corresponding statistical analysis. Both datasets comprise parameters of EV batteries under real-world conditions.

Dataset A provides long-term charging data from 20 commercial EVs with identical battery systems, each monitored over approximately 29 months.[20] The data were collected during charging via CAN communication at regular intervals and captured key patterns relevant to real-world battery health evaluation. The metric for battery usage is the length of time in service, which is reasonable given that commercial EVs are in continuous operation.

Dataset B offers a large-scale time-series capacity data of 191 EVs, including over 1.2 million charging sessions from vehicles across three manufacturers.[21] Each session records multiple charging-related parameters at fixed intervals, including voltage, current, temperature, capacity, and estimated SOC. The dataset is designed to facilitate deep learning research on charging behavior, battery degradation, safety, and energy management in real-world settings. The usage metric in this dataset is the odometer reading.

In both datasets, each EV has an average of over 2,000 data points of battery capacity. An overview of these datasets is shown in Table 1.

Table 1: Overview of two EV battery datasets: dataset A from 20 commercial EVs monitored over approximately 29 months, dataset B from 191 EVs with over 1.2 million charging sessions.

Dataset	#EVs	#Avg. Points per EV	#Total Points	Usage Metric
Dataset A[20]	20	~2,696	53,927	Time in service (day)
Dataset B[21]	191	~3,068	585,922	mileage (km)

Let C denote the capacity of EV batteries, and x denote the usage metric. A two-step process is used to investigate the relationship between C and x. First, the correlation coefficient between C and x is calculated to check if their correlation is indeed negative, as intuitively expected. The linear regression model is then estimated:

$$C = \beta x + \alpha \tag{1}$$

Table 2: Correlation and linear regression analysis between battery capacity C and usage metrics x. A strong negative correlation is found in both datasets. The linear regression models are $C = -2.228 \times 10^{-2} * \textit{Time in Service} + 132.573$ for Dataset A, and $C = -2.554 \times 10^{-5} * \textit{Mileage} + 43.308$ for Dataset B.

Dataset	Correlation between C and x	Parameter	Estimate	95% Confidence Interval
Dataset A	–0.709	β	-2.228×10^{-2} *	[-2.247×10^{-2}, -2.209×10^{-2}]
		α	132.573 *	[132.475, 132.671]
Dataset B	–0.695	β	-2.554×10^{-5} *	[-2.561×10^{-5}, -2.548×10^{-5}]
		α	43.308 *	[43.301, 43.315]

Notes: * indicates significance at the $p < 0.001$ level.

The following observations are made based on the results shown in Table 2:

(1) C and x exhibit a strong negative correlation in both datasets (–0.709 for Dataset A and –0.695 for Dataset B), consistent with the well-known fact that battery capacity decreases with increased usage.

(2) Recall the substantial differences between the two datasets in terms of battery types, EV models, data collection conditions, and usage metrics. Note that both datasets result in correlation coefficients close to –0.7, which indicates that the rate of battery capacity degradation with usage is consistent.

(3) Data points contain much noise, highlighting the difficulty of accurately predicting battery capacity at the individual level under real-world conditions. The considerable noise may be attributed to complex environmental factors that lead to a wide range of data fluctuations. This suggests that a large number of samples (data points) is necessary to effectively mitigate the impact of noise on parameter estimation.

(4) The 95% confidence intervals for the parameters are very narrow, indicating a low degree of uncertainty in the parameter estimates. We also show the curve fitting of Datasets A and B in Figures 1 and 2, respectively.

DOI: 10.36838/v8i2.50

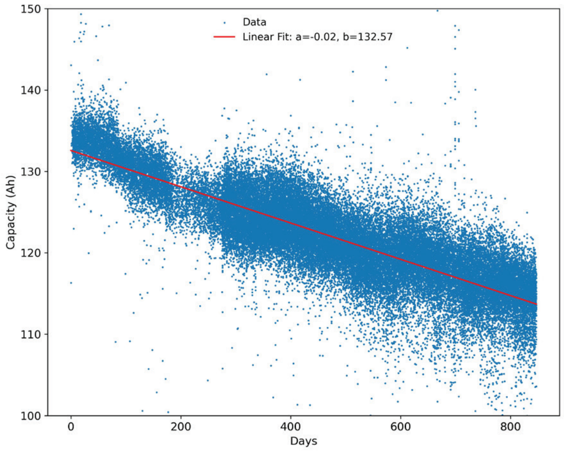


Figure 1: Battery capacity vs number of days in service in Dataset A: scatter plot in blue and linear regression line C = -2.228 x 10^{-2} $*$ *Time in Service* + 132.573 in red.

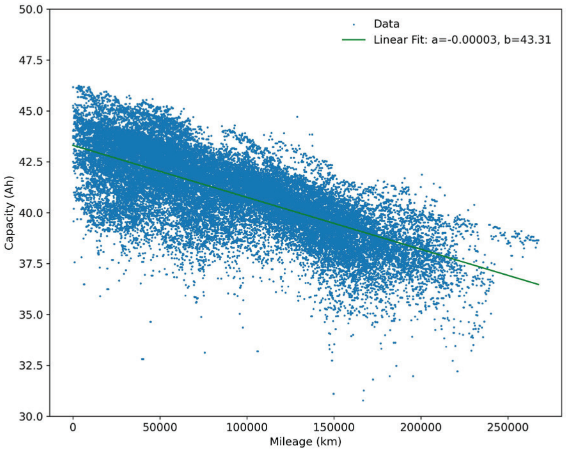


Figure 2: Battery capacity vs mileage in Dataset B: scatter plot in blue and linear regression line C = -2.554 x 10^{-5} $*$ *Mileage* + 43.308 in dark green.

It is observed from Figures 1 and 2 that the capacity of EV batteries decreases linearly with increased usage. These linear models form the basis for our discussion on the cascade utilization flow model in the next section.

Optimization of Cascade Utilization for Social Welfare:

This subsection explores how the government could manage the cascade utilization to maximize social welfare. The following notations are used in subsequent discussions.

$D_t^{(EV)}$ and $D_t^{(Non\text{-}EV)}$: demand for batteries by the EV and non-EV population at time step t, respectively.

$B_t^{(EV)}$ and $B_t^{(Non\text{-}EV)}$: the number of new batteries that need to be produced for the EV and non-EV population at time step t, respectively.

$I_t^{(EV)}$ and $I_t^{(Non\text{-}EV)}$: the number of batteries in the EV and non-EV population at time step t, respectively.

$I_{t,c}^{(EV)}$ and $I_{t,c}^{(Non\text{-}EV)}$: the number of batteries with capacity c in the EV and non-EV population at time step t, respectively.

$r_t^{(EV)}$ and $r_t^{(Non\text{-}EV)}$: the number of batteries to be recycled from the EV and non-EV population at time step t, respectively.

s_t: the number of batteries cascaded from the EV population to the non-EV population at time step t.

Figure 3 illustrates the flow of EV and non-EV batteries. The demand for EV batteries ($D_t^{(EV)}$), is satisfied by batteries that are currently in the EV population and the number of new batteries produced ($B_t^{(EV)}$). Due to capacity degradation, batteries will no longer meet the capacity requirements of the EV population after a period of usage. Some ($r_t^{(EV)}$) need to be directly recycled, while others (s_t) still hold value for cascade utilization in the non-EV population. Therefore, demand for non-EV batteries ($D_t^{(Non\text{-}EV)}$), is satisfied by current batteries in the non-EV population, new production for non-EV usage ($B_t^{(Non\text{-}EV)}$), and batteries cascaded from the EV population (s_t). Batteries in the non-EV population also degrade, and some need to be recycled ($r_t^{(Non\text{-}EV)}$). By knowing the state at each timestamp, the evolution of the battery population can be captured starting from t = 0 onward.

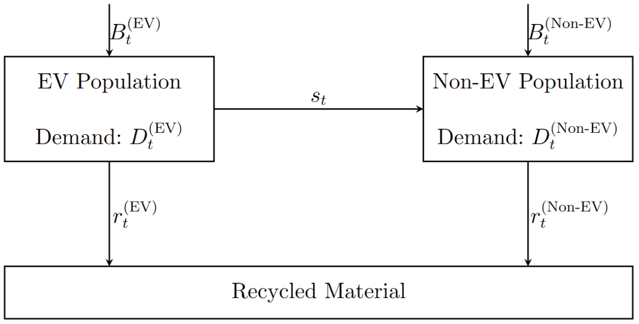

$B_t^{(\mathrm{EV})}$
$B_t^{(\mathrm{Non\text{-}EV})}$
EV Population
Demand: $D_t^{(\mathrm{EV})}$
s_t
Non-EV Population
Demand: $D_t^{(\mathrm{Non\text{-}EV})}$
$r_t^{(\mathrm{EV})}$
$r_t^{(\mathrm{Non\text{-}EV})}$
Recycled Material

Figure 3: Flow of batteries in the market at time step t. The demand for EV batteries ($D_t^{(EV)}$) is satisfied by batteries that are currently in the EV population and the number of new batteries produced ($D_t^{(EV)}$). Demand for non-EV batteries ($D_t^{(Non\text{-}EV)}$) is satisfied by current batteries in the non-EV population, new production for non-EV usage ($B_t^{(Non\text{-}EV)}$), and batteries cascade utilized from the EV population (s_t). Some batteries ($r_t^{(EV)}$, $r_t^{(Non\text{-}EV)}$) are recycled.

In the process of cascade utilization of EV batteries, government intervention is often necessary to maximize social welfare. Government policies and regulations can provide essential guidelines for the proper management of battery resources, ensure environmental protection, and promote sustainable economic development. A better understanding of the dynamics between cascade utilization and recycling of batteries will guide more effective government policies.

Let c_0 denote the initial capacity of a battery, and c_R the recycling threshold for EV batteries. Similarly, c_S is the threshold for cascade utilization, and $c_S > c_R$. For all EV batteries with a capacity of c_S, we stipulate that no more than a proportion q of them will be cascade utilized, while the rest will continue to be used in the EV market until they are recycled. In this process, the initial capacity c_0 and recycle capacity c_R are determined by the characteristics of batteries, while the cascade capacity c_S and cascade ratio q can be adjusted by the government. These standards can directly affect the flow of batteries, including production, supply, utilization, and recycling. Therefore, we want to explore how the standards should be developed to enhance the overall societal benefits.

Battery Dynamics in the EV Population:

The batteries used in the EV population gradually degrade in daily driving and end up in the non-EV population or are recycled. To accurately model the battery flow in the EV population, let δc be the capacity degraded during one time step, and $I_{t,c}{}^{(EV)}$ be the number of batteries with capacity c at time step t. At each time step t, the number of batteries with capacity c_0 in the EV population is equal to the number of batteries produced for the EV population at time step t:

$$I_{t,c_0}^{(EV)} = B_t^{(EV)} \tag{2}$$

The batteries with capacity c_S + δc in the EV population will degrade to capacity c_S, which is the capacity threshold for cascade utilization. Considering that not all batteries can be collected for cascade utilization, we assume only a proportion q of them can be transferred into the non-EV population. Furthermore, the demand for new non-EV batteries also restricts the number of batteries transferred. Therefore, the batteries transferred from the EV population to the non-EV population can be expressed as:

$$s_t = \min\{I_{t,c_S+\delta c}^{(EV)} \cdot q, \max\{0, D_t^{(Non-EV)} - I_t^{(Non-EV)}\}\} \tag{3}$$

and the number of batteries with capacity c_S in the EV population at time step t + 1 is equal to the number of batteries that are not transferred into the non-EV population,

$$I_{t+1,c_S}^{(EV)} = I_{t,c_S+\delta c}^{(EV)} - s_t \tag{4}$$

When batteries with capacity c_R + δc degraded to c_R, they are forced to be recycled,

$$r_t^{(EV)} = I_{t,c_R+\delta c}^{(EV)} \tag{5}$$

Then, the number of batteries with capacity c_R in the EV population at time step t becomes 0,

$$I_{t,c_R}^{(EV)} = 0 \tag{6}$$

For batteries in other capacity ranges, the number of batteries with capacity c in the EV population at time t + 1 is equal to the number of batteries with capacity c + δc in EV users at time step t. To be specific, this applies to the capacity ranges $c_S + \delta c \le c \le c_0 - \delta c$ and $c_R + \delta c \le c \le c_S - \delta c$. This degradation process can be expressed as:

$$I_{t+1,c}^{(EV)} = I_{t,c+\delta c}^{(EV)}, \text{for } c_S + \delta c \le c \le c_0 - \delta c \text{ and } c_R + \delta c \le c \le c_S - \delta c \tag{7}$$

At time step t, the total number of batteries in EV users $I_t{}^{(EV)}$ is the sum of the batteries with capacities ranging from c_R to c_0:

$$I_t^{(EV)} = \sum_{c=c_R}^{c_0} I_{t,c}^{(EV)} \tag{8}$$

Then for each time step t, the total number of batteries in EV users at time step t should be no less than the battery demand of the EV population at time step t:

$$I_t^{(EV)} \ge D_t^{(EV)} \tag{9}$$

Equivalently, the number of batteries produced for the EV population can be expressed as:

$$B_t^{(EV)} = \max\{D_t^{(EV)} - I_t^{(EV)}, 0\} \tag{10}$$

Battery Dynamics in the Non-EV Population:

The batteries used in the non-EV population are either batteries produced for non-EV usage or are from the EV population. Non-EV batteries can be used for energy storage and power supply for communication base stations, power stations, and in other commercial settings. These batteries will also gradually degrade over time and end up being recycled. We assume that they follow the same degradation pattern as the EV batteries. Similar to the modeling of the EV population, let $I_t{}^{(Non\text{-}EV)}$ be the number of batteries in the non-EV population with capacity c at time step t, and δc be the capacity degraded within one time step. The number of batteries with capacity c_0 in the non-EV population at time t is equal to the number of batteries produced for the non-EV population at time step t,

$$I_{t,c_0}^{(Non-EV)} = B_t^{(Non-EV)} \tag{11}$$

The number of batteries with capacity c_S at time step t + 1 is equal to the number of batteries with capacity c_S + δc at time step t plus the number of batteries transferred from the EV population s_t,

$$I_{t+1,c_S}^{(Non-EV)} = I_{t,c_S+\delta c}^{(Non-EV)} + s_t \tag{12}$$

When batteries with capacity c_R + δc degraded to the standard non-EV battery recycle capacity c_R, they are recycled,

$$r_t^{(Non-EV)} = I_{t,c_R+\delta c}^{(Non-EV)} \tag{13}$$

The number of batteries with capacity c_R at time step t is 0, indicating that all batteries of this capacity are recycled,

$$I_{t,c_R}^{(Non-EV)} = 0 \tag{14}$$

Batteries of other capacities follow a normal degradation, which means that the number of batteries with capacity c at time t + 1 is equal to the number of batteries with capacity c + δc at time t,

$$I_{t+1,c}^{(Non-EV)} = I_{t,c+\delta c}^{(Non-EV)}, \text{for } c_S + \delta c \le c \le c_0 - \delta c \text{ and } c_R + \delta c \le c \le c_S - \delta c \tag{15}$$

At time step t, the total number of batteries $I_t{}^{(Non\text{-}EV)}$ is the sum of the batteries with capacities ranging from c_R to c_0,

$$I_t^{(Non-EV)} = \sum_{c=c_R}^{c_0} I_{t,c}^{(Non-EV)} \tag{16}$$

DOI: 10.36838/v8i2.50

The total number of batteries at time step *t* should be able to cover the demand for batteries at time step *t*,

$$I_t^{(Non-EV)} \geq D_t^{(Non-EV)} \tag{17}$$

Equivalently, the batteries to be produced for the non-EV population can be expressed as

$$B_t^{(Non-EV)} = \max\{D_t^{(Non-EV)} - I_t^{(Non-EV)} - s_t, 0\} \tag{18}$$

Social Welfare Modeling:

With the formulation of the battery dynamics in the EV and non-EV population, we further formulate the social welfare of this process. In this context, social welfare is the benefits that battery usage brings to society. For example, using EV batteries with higher average capacity can improve the overall efficiency of EV usage, thereby enhancing the efficiency of the entire transportation system and contributing to greater social welfare. Conversely, lower capacity reduces efficiency, leading to lower social welfare. We assume the social welfare brought by the batteries used in the EV and non-EV populations is related to the mean capacity in each population. The mean capacities can be calculated as:

$$\boldsymbol{c_t^{(EV)}} = \sum_{\boldsymbol{c=c_R}}^{\boldsymbol{c_0}} \boldsymbol{I_{t,c}^{(EV)}}\, c/\boldsymbol{I_t^{(EV)}} \tag{19}$$

$$c_t^{(Non-EV)} = \sum_{c=c_R}^{c_0} I_{t,c}^{(Non-EV)}\, c/I_t^{(Non-EV)} \tag{20}$$

We use a utility function $f^{(EV)}$ to characterize the social welfare within the EV population. For EV owners, when the degradation starts from a brand-new battery, the major effect is the gradual reduction in range. However, as battery capacity continues to degrade, some other issues become more apparent, including deterioration in acceleration and braking performance, slowing down of charging speeds, and a greater failure rate of the vehicle's information systems.[22] Therefore, we propose to employ a non-linear utility function $f^{(EV)}$ of mean EV capacity to measure the unit social welfare of the EV population:

$$f^{(EV)}(c) = \begin{cases} k_1(c - c_R), \text{if } c_R \leq c < c^{EV} \\ k_1(c^{EV} - c_R) + k_2(c - c^{EV}), \text{if } c^{EV} \leq c \leq c_0 \end{cases} \tag{21}$$

where $k_2 < k_1$ are the slopes for the piece-wise linear utility function, and c^{EV} is the threshold where the slope changes. Similarly, the unit social welfare of the non-EV population can also be measured by a non-linear function:

$$f^{(Non-EV)}(c) = \begin{cases} k_3(c - c_R), \text{if } c_R \leq c < c^{Non-EV} \\ k_3(c^{Non-EV} - c_R) + k_4(c - c^{Non-EV}), \text{if } c^{Non-EV} \leq c \leq c_0 \end{cases} \tag{22}$$

where $k_4 < k_3$ are slopes of the utility function $f^{(Non-EV)}$, and c^{Non-EV} is the threshold where the slope changes. The non-EV population often has lower requirements on batteries; therefore, it is reasonable to assume that $c^{Non-EV} < c^{EV}$ and $k_4(c^{Non-EV} - c_R) \geq k_1(c^{EV} - c_R)$.

Based on the unit social welfare and the battery ownership in the EV and non-EV population, their total social welfare can be respectively expressed as:

$$\boldsymbol{u_t^{(EV)}} = \boldsymbol{f^{(EV)}}\left(\boldsymbol{c_t^{(EV)}}\right) \cdot \boldsymbol{I_t^{(EV)}} \tag{23}$$

$$u_t^{(Non-EV)} = f^{(Non-EV)}\left(c_t^{(Non-EV)}\right) \cdot I_t^{(Non-EV)} \tag{24}$$

We assume that new batteries are manufactured with unit cost c_m. With the number of batteries produced for both EV and non-EV populations, the related social welfare at time step *t* is $- c_m\,(B_t^{(EV)} + B_t^{(Non-EV)})$.

To make EV batteries available for non-EV utilization, the following costs must be considered.[23] First, the batteries need to be dismantled, which requires labor and equipment costs. Before entering the non-EV population, batteries need to be inspected to determine if they meet the non-EV utilization standards based on their remaining life and performance. Batteries may also need to be repackaged or remanufactured before entering the cascade utilization market to ensure safe transportation and storage. We denote the unit transfer cost by k_s, which covers all the costs mentioned above, and the related social welfare in time step t can be expressed as $- k_s\, s_t$.

For recycling end-of-life batteries, there are also several costs to be considered. The end-of-life batteries often need to be centralized for further processing because of the potential hazards and pollution that can result from the process. Therefore, the logistics cost incurred during battery collection and transportation is not negligible. The collected end-of-life batteries also need to be sorted and pre-processed, which means labor costs, the cost of sorting equipment, and the cost of preliminary discharge and disassembly of batteries. Then these batteries are processed using chemical, mechanical, and thermal treatment methods, which incur costs for equipment usage, chemical reagents, energy consumption, and labor. These processes are often accompanied by useless or even harmful by-products, like waste liquids, residues, and gases. The cost of handling those hazardous substances should also be considered. In our formulation, we denote the unit battery recycle cost by kr to account for all the costs mentioned above in the recycling process. Since the batteries can be recycled from the EV and non-EV population, the related social welfare at time step t is expressed as $- k_r\,(r_t^{(EV)} + r_t^{(Non-EV)})$.

Based on the above formulation of the EV and non-EV battery dynamics and social welfare, the decision-making of the government can be advised by the following optimization problem:

$$\max_{c_S\,,q} \frac{1}{T} \sum_t u_t^{(EV)} + u_t^{(Non-EV)} - k_s s_t - k_r\left(r_t^{(EV)} + r_t^{(Non-EV)}\right) - c_m\left(B_t^{(EV)} + B_t^{(Non-EV)}\right) \tag{25}$$

Subject to: EV battery dynamics in Eqs. (2-8), and (10),
Non-EV battery dynamics in Eqs. (11-16), and (18),
Capacity calculation in Eqs. (19) and (20).

Social welfare ($u_t^{(EV)}$, $u_t^{(Non-EV)}$) is calculated using Eqs (23) and (24). Recall that the parameter *q* is the maximum cascade ratio of batteries, and the actual number of batteries that can be transferred to the non-EV population is also restricted by the demand of the non-EV population. Intuitively speaking, a lower q limits the cascade utilization of batteries. Therefore, if a battery can bring more social utility after entering the non-EV population, the total social welfare would increase with q. On the other hand, if a battery can bring more social utility when staying in the EV population, the total social welfare may decrease with a larger q, because a smaller q can keep more batteries in the EV population and result in more social welfare.

The effect of c_S is much more complex. From the perspective of the life cycle of an individual battery, the manufacturing cost and the recycling cost are fixed and not influenced by c_S. If the battery stays in the EV population until recycled, social welfare is not influenced by c_S. If a battery is transferred to the non-EV population at capacity c_S, a higher c_S generates greater social welfare in the non-EV population compared to the EV population, as indicated by our formulation of the functions $f^{(EV)}(c)$ and $f^{(Non\text{-}EV)}(c)$, along with the associated parameter requirements. If the difference in social welfare between the EV and non-EV population exceeds the transfer cost ks, the more batteries transferred, the more social welfare is achieved. A higher c_S means a battery can serve for a longer time in the non-EV population until recycled, which lowers the demand of the non-EV population, and results in a reduction in the volume of batteries transferred. In addition, a higher c_S may also accelerate the replacement of batteries in the EV population, which incurs more cost for the production of new batteries. These effects we discussed are also highly dependent on the parameter settings and the real-world demand and capacity distribution. Therefore, the overall effect of c_S is hard to predict.

Model Assumptions and Limitations:

Before we discuss computational results, it should be noted that several assumptions are made in this study. First, the battery degradation model assumes linear capacity degradation, which may differ from actual nonlinear degradation patterns influenced by varying operational conditions. Second, the battery population is treated as homogeneous despite variations in battery chemistries and usage scenarios. Finally, our predictions rely on logistic growth modeling of market demand, which may not fully capture market uncertainties or disruptive technological changes. These limitations should be considered when interpreting the results.

■ Result and Discussion

Estimated Growth of EV and Non-EV Batteries:

We collected data on EV ownership in the Chinese market from 2017 to 2023.[24] Based on this data, we estimate the future growth of the EV market. Three models are selected for this estimation: the exponential model, the quadratic function model, and the logistic growth model, with mean squared error (MSE) used as the evaluation metric. Figure 4 shows the fitting performance of these three models, with time in the horizontal axis, where a real-valued interval $[i, i+1)$ represents Year i, and the number of EVs in the vertical axis.

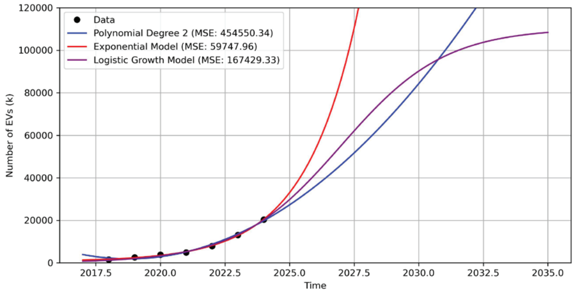


Figure 4: Estimate of EV growth in China based on data from 2017-2023. Three models - exponential, quadratic function, and logistic growth – are estimated. Mean squared error (MSE) for each model is included as a measure of fitness. Since future growth cannot increase indefinitely as predicted in the exponential model and the quadratic function model, the logistic growth model is selected as the preferred model.

Although the exponential model provides the best fit for the 2017-2023 data, it should be noted that using it to predict future growth could be problematic since the number of EVs cannot grow indefinitely as predicted in the exponential model and the quadratic function model. Furthermore, the logistic growth model exhibits low fitting errors, indicating a decent fit. It is also noteworthy that, according to our proposed model, EV ownership will reach approximately 100 million by the end of 2030. This prediction aligns with the forecast made by leading experts, which underscores the validity and reliability of our proposed model.[25]

According to EVChina, the most common cascade uses of EV batteries are for energy storage in communication base stations, renewable energy storage, and public facility energy storage.[26] We aggregate the demand data for these three applications from 2018 to 2022 to obtain the overall demand data for the cascade use. Based on the data, we fit a logistic growth model for the demand of non-EV batteries (note that this amount is calculated based on the capacity of EV batteries). Figure 5 shows the fitting result. It can be observed that the scale of electricity usage for these non-EV batteries experiences a period of rapid growth, followed by a slowdown in growth, and eventually stabilizes around 2035.

The estimated numbers of EV and non-EV batteries are used as the initial conditions, $D_1^{(EV)}$ and $D_1^{(Non\text{-}EV)}$, in the cascade flow model.

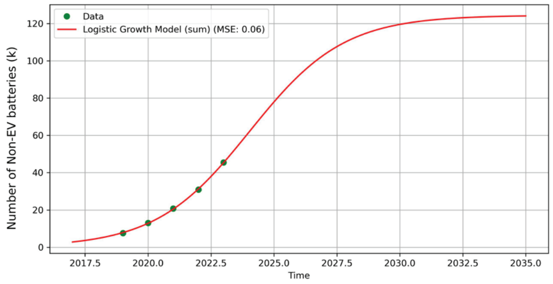


Figure 5: Logistic growth curve for non-EV batteries in China. The logistic growth model is estimated using data from 2018-2022 that cover energy storage in communication base stations, renewable storage, and public facility storage.

DOI: 10.36838/v8i2.50

Macro Perspective on Cascade Utilization Flow Model:

The evolution of battery population is simulated using the flow model, including EV battery dynamics in Eqs. (2-8), and (10), non-EV battery dynamics in Eqs. (11-16), and (18), and capacity calculation in Eqs. (19) and (20). The corresponding social welfare ($u_t^{(EV)}$, $u_t^{(Non\text{-}EV)}$) is calculated using Eqs. (23-25).

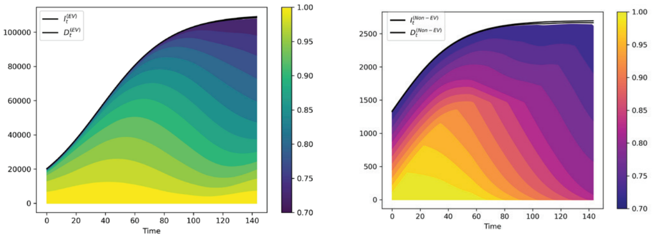
The parameter settings for the simulation study are listed in Table 3. More specifically, the parameters for time are based on our defined study period, parameters for battery degradation are from the empirical analysis discussed in the Electric Vehicle Battery Capacity Degradation Model section of this paper, cost-related parameters are adopted using a normalization approach based on the conceptual frameworks in related studies,[18,19] and the social utility parameters are based on the key contribution of our model, designed to capture the proposed non-linear welfare effects of battery performance.[18]

Table 3: Parameter settings for the cascade utilization flow model. The battery population evolution is simulated for 12 years (144 time steps; each time step is a month).

Parameters	Values
T	144 (12 years)
k_s	0.3
k_r	0.01
c_m	1
c_0	1
δc	1 / 480
c^{EV}	0.8
k_1	8
k_2	1
$c^{Non\text{-}EV}$	0.8
k_3	40
k_4	5 / 7

Figure 6 shows the total number of batteries, as well as the changes in battery capacity structure. We partition the capacity value of [0.7, 1] into 12 equal intervals and mark intervals with different colors. For EV batteries, the ratio of batteries of high capacity (capacity 0.85 ~ 1) will first increase and then decrease. This is because the number of EV batteries will initially undergo a rapid growth phase, during which a large number of new batteries with high capacity will enter the population, increasing their proportion. As growth slows, the number of new batteries entering the population each period will decrease. Additionally, the capacity of batteries from the previous high-growth phase will gradually degrade, leading to an increase in the proportion of low-capacity batteries (capacity 0.7 ~ 0.85).

In the initial phase (time step 1 to 40), the proportion of high-density batteries in non-EVs is rising. This is because there are too few EV batteries available for cascade utilization to meet the non-EV demand at this stage. Consequently, additional new batteries need to be produced for non-EV applications. These high-capacity new batteries increase their proportion in the population. After this initial phase, the overall capacity within the non-EV population rapidly declines. By the end of the simulation, almost all non-EV batteries originate from the cascade utilization of EV batteries. This shift is due to the rapid growth in EV batteries, which significantly increases the number of EV batteries available for cascade utilization, adequately meeting the non-EV electricity demand.



(a)EV batteries

(b) Non - EV batteries

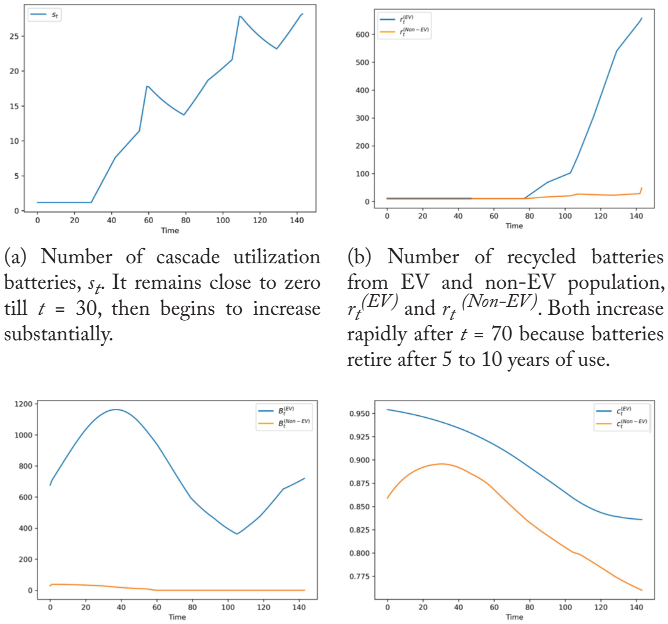
Figure 6: Change in EV and non-EV battery populations and their capacity distributions over time. The capacity range [0.7, 1] is divided into 12 equal intervals, represented by different colors. In the EV battery population, the proportion of high capacity (capacity > 0.85) batteries will increase initially and then decrease, proportion of low capacity (capacity between 0.7 and 0.85) batteries will increase due to capacity degradation over time. In the non-EV population, after an initial phase with a high proportion of high-capacity batteries, the overall capacity rapidly declines.

Deeper Analysis of Cascade Utilization Dynamics:

In Figure 7, other key variables in the model are presented to better understand the dynamics of cascade utilization.

- The number of EV batteries for cascade utilization, st, remains close to zero during the period from $t = 0$ to $t = 30$, after which it begins to increase. This initial phase sees a scarcity of EV batteries suitable for cascade utilization. However, as the scale of EV batteries rapidly grows, each period witnesses a substantial number of EV batteries degrading to the capacity threshold c_S, making them available for cascade utilization and leading to the subsequent increase in s_t.
- Both $r_t^{(EV)}$ and $r_t^{(Non\text{-}EV)}$, the numbers of batteries recycled from the EV and non-EV populations, have a rapid increase after t = 70. This is due to both market demands experiencing rapid growth phases, with these batteries gradually retiring after 5 to 10 years of use, leading to a significant increase in r_t. This also warns us that if we cannot effectively manage the impact of retired batteries, our environment will be severely polluted by the chemical elements contained in these retired batteries.
- The production of EV batteries each period $B_t^{(EV)}$ exhibits three phases: an initial increase, followed by a decrease, and then another increase. The trends in the first and second phases are due to the rapid initial growth rate of required EV batteries, which then slows down. The increase in the third phase is attributed to the large number of batteries retiring from earlier periods, necessitating the production of new EV batteries to meet this demand.
- The average capacity of non-EV batteries initially increases briefly and then continues to decline. This is because, in the early stages, non-EV batteries require new batteries to meet rapidly increasing demand. As the scale of EVs expands and

the number of EV batteries available for cascade utilization increases, the non-EV battery demand can be adequately met by these cascade-utilized EV batteries, leading to a continuous decrease in average capacity.



(a) Number of cascade utilization batteries, s_t. It remains close to zero till t = 30, then begins to increase substantially.

(b) Number of recycled batteries from EV and non-EV population, $r_t^{(EV)}$ and $r_t^{(Non-EV)}$. Both increase rapidly after t = 70 because batteries retire after 5 to 10 years of use.

(c) Battery production for EV and non-EV, $B_t^{(EV)}$ and $B_t^{(Non-EV)}$. $B_t^{(EV)}$ experiences a rapid initial increase to meet the demand, followed by a decrease, and then another increase when the initial batteries retire and more new ones need to be produced.

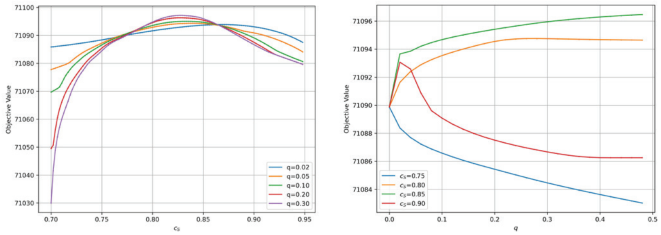
(d) Average capacity of EV and non-EV population, $c_t^{(EV)}$ and $c_t^{(Non-EV)}$. $c_t^{(Non-EV)}$ increases briefly and then continues to decline. This is because the initial demand for non-EV is satisfied by new batteries but later the demand can be adequately met by cascade utilized batteries.

Figure 7: Key variables in the cascade utilization flow model.

Sensitivity Analysis of Key Parameters:

To maximize the objective function of social welfare, we need to understand at what capacity level (c_S) and in what proportion (q) EV batteries should be cascade utilized. We analyze the changes in the objective function under two scenarios: (1) varying c_S while keeping q constant, and (2) varying q while keeping c_S constant. Figure 8 shows the results from the two scenarios.

- Varying c_S with fixed q. At different levels of q, we vary c_S from 0.7 to 0.95, covering a large range of battery capacity. The curves in Figure 8(a) consistently exhibit an increase followed by a decrease as cS increases. This trend is due to the trade-off between EV and non-EV batteries. If c_S is low, the number of EV batteries that need to be produced each period is reduced, which lowers the cost of producing new EV batteries. However, the average capacity of batteries available for cascade utilization will also be lower, resulting in lower social welfare for the non-EV sector. Conversely, if c_S is high, the average capacity of EV batteries will increase social welfare in the EV sector, but more EV batteries will need to be produced. The batteries available for cascade utilization will have a higher average capacity, thereby increasing the social welfare in the non-EV sector. These factors interact, ultimately leading to an objective function curve that initially increases and then decreases. This also implies that, for a fixed level of q, there exists an optimal value for c_S somewhere between c_R and c_0.
- Varying q with fixed c_S. Intuitively, we believe that a higher proportion of batteries available for cascade utilization can increase the objective function. However, our results as shown in Figure 8(b) indicate that this intuition only holds true when the value of c_S is appropriate. If the value of c_S is too low (that is, close to the mandatory recycling level c_R), then the average capacity of the cascade utilized batteries will be low. Therefore, increasing q will result in the non-EV population being flooded with nearly obsolete batteries, causing a decline in the objective function. On the other hand, if the value of c_S is too high, then batteries are utilized for cascade applications early in their life cycle, and significantly more EV batteries will need to be produced each period, again leading to a decline in the objective function. Only when the value of c_S is appropriate -- the battery performance is no longer sufficient to meet the requirements of EV usage but can still satisfy the needs of cascade utilization -- will increasing q lead to a continuous increase in the objective function.

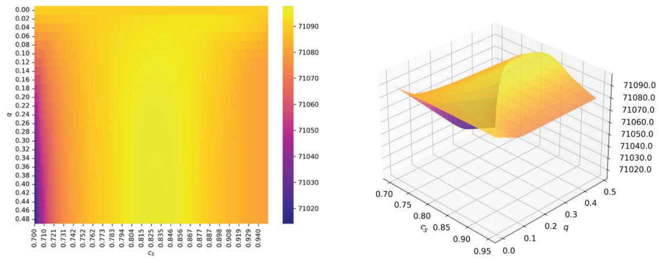


(a) Social utility as a function of c_S, capacity level for cascade use. For a fixed level of q, there exists an optimal value for c_S that maximizes social welfare.

(b) Social utility as a function of q, proportion of EV batteries that could be utilized for cascade use.

Figure 8: Social utility as a function of c_S and q.

To visualize how the objective function is affected by both c_S and q simultaneously, we show the variations in the objective function values on a parameter grid, as shown in Figure 9. The parameter grid is spanned by q = [0, 0.5) with interval 0.02 and c_S = [0.7, 0.95) with interval δc, and the maximum value is attained when q = 0.48 and c_S = 0.829. It is conceivable that if we continue to increase the value of q, the objective may still have minor increases, because increasing q means increasing the number of batteries available for cascade utilization, which provides the potential for higher social welfare.



(a) 2D Heatmap of objective values on parameter grids.

(b) 3D Heatmap of objective values on parameter grids.

Figure 9: Heatmap of objective values on parameter grids of c_S and q. Parameter q ranges from 0 to 0.5, and parameter c_S from 0.7 to 0.95. The maximum objective value is attained when q = 0.48 and c_S = 0.829.

Conclusion

This paper presents a comprehensive analysis of the lifecycle management of EV batteries, emphasizing the dual strategies of recycling and cascade utilization. By leveraging extensive real-world data, we developed a model that predicts battery lifespan and performance, providing a robust foundation for policy and strategic decisions. The alignment of our model's predictions from a macro market perspective with expert forecasts[23] underscores the validity and reliability of the model and demonstrates its practical applicability in real-world scenarios. This model also forms the basis for subsequent analyses.

Sensitivity analysis of key parameters is conducted to identify the most impactful factors on system performance. This analysis reveals the importance of optimizing the cascade ratio and recycling efficiency to maximize social welfare. Policymakers should consider these findings when formatting regulations and incentives to ensure they address the most critical aspects of battery lifecycle management.

The following recommendations, derived from numerical findings and analysis, provide a roadmap for policymakers to enhance the sustainability and economic viability of EV battery lifecycle management.

1. Enhancing data collection and sharing: Governments should promote the establishment of comprehensive databases for battery usage and degradation data. This would improve model accuracy and enable better lifecycle management of EV batteries.

2. Establishing robust recycling standards: Implementing strict recycling standards can ensure that retired batteries are processed in an environmentally friendly manner, minimizing hazardous waste and promoting the recovery of valuable materials.

3. Supporting technological innovation: Investing in research and development for advanced battery technologies and recycling processes can drive innovation, improve recycling efficiency, and reduce costs. This includes supporting the development of more efficient battery degradation models.

4. Developing infrastructure for battery management: Building robust infrastructure for the collection, transportation, and processing of batteries is essential. This includes creating facilities for recycling and cascade utilization to ensure efficient handling of retired batteries.

5. Monitoring the advancements in battery technology: Adjustments to policy decisions regarding cascading utilization should be made in response to changes in the patterns of battery capacity degradation, based on the prevailing circumstances.

In conclusion, this study provides insights for policymakers and industry stakeholders and presents a path forward for improving the sustainability and economic viability of EV batteries. Future research should continue to refine and expand to more complicated battery degradation models, to incorporate emerging data, and to explore new strategies to further enhance EV battery management. By doing so, we can ensure that the rapid growth of the EV industry contributes positively to both economic development and environmental sustainability.

Acknowledgments

I would like to thank my research mentor, Dr. Dazhou Lei, for his patience, guidance, and support throughout the research process.

References

1. Canalys. Global EV Market Forecasted to Reach 17.5 Million Units with Solid Growth of 27% in 2024. 2024. https://www.canalys.com/newsroom/global-ev-market-2024.
2. International Energy Agency. Global EV Outlook 2025. 2024.
3. Zhao, Y.; Wang, Z.; Shen, Z.-J. M.; Sun, F. Assessment of Battery Utilization and Energy Consumption in the Large-Scale Development of Urban Electric Vehicles. *Proceedings of the National Academy of Sciences* **2021**, *118* (17), e2017318118.
4. Wu, Y.; Yang, L.; Tian, X.; Li, Y.; Zuo, T. Temporal and Spatial Analysis for End-of-Life Power Batteries from Electric Vehicles in China. *Resour Conserv Recycl* **2020**, *155*, 104651.
5. Wang, S. Multi-Angle Analysis of Electric Vehicles Battery Recycling and Utilization. In *IOP conference series: Earth and environmental science*; **2022**; Vol. 1011, p 12027.
6. Xie, Y.; Yu, H.; Li, C. D. Research on Systems Engineering of Recycling EV Battery. In *Proceedings of the 2nd International Conference on Advances in Mechanical Engineering and Industrial Informatics (AMEII 2016), Hangzhou, China*; 2016; pp 9–10.
7. Ahmadi, L.; Young, S. B.; Fowler, M.; Fraser, R. A.; Achachlouei, M. A. A Cascaded Life Cycle: Reuse of Electric Vehicle Lithium-Ion Battery Packs in Energy Storage Systems. *Int J Life Cycle Assess* **2017**, 22, 111–124.
8. Zhu, C.; Xu, J.; Liu, K.; Li, X. Feasibility Analysis of Transportation Battery Second Life Used in Backup Power for Communication Base Station. In *2017 IEEE Transportation Electrification Conference and Expo, Asia-Pacific (ITEC Asia-Pacific)*, 2017; pp 1–4.
9. Valant, C.; Gaustad, G.; Nenadic, N. Characterizing Large-Scale, Electric-Vehicle Lithium Ion Transportation Batteries for Secondary Uses in Grid Applications. *Batteries* **2019**, *5* (1), 8.
10. Richa, K.; Babbitt, C. W.; Nenadic, N. G.; Gaustad, G. Environmental Trade-Offs across Cascading Lithium-Ion Battery Life Cycles. *Int J Life Cycle Assess* **2017**, *22*, 66–81.
11. China Electricity Council. 2022 Annual Statistics of the Electrochemical Energy Storage Power Station Industry. March 2023.
12. Edge, J. S.; O'Kane, S.; Prosser, R.; Kirkaldy, N. D.; Patel, A. N.; Hales, A.; Ghosh, A.; Ai, W.; Chen, J.; Yang, J.; others. Lithium Ion Battery Degradation: What You Need to Know. *Physical Chemistry Chemical Physics* **2021**, *23* (14), 8200–8221.
13. Luo, G.; Zhang, Y.; Tang, A. Capacity Degradation and Aging Mechanisms Evolution of Lithium-Ion Batteries under Different Operation Conditions. *Energies (Basel)* **2023**, *16* (10).
14. Zhang, Y.; Tang, Q.; Zhang, Y.; Wang, J.; Stimming, U.; Lee, A. A. Identifying Degradation Patterns of Lithium Ion Batteries from Impedance Spectroscopy Using Machine Learning. *Nat Commun* **2020**, *11* (1), 1706.
15. Huang, H.; Bian, C.; Wu, M.; An, D.; Yang, S. A Novel Integrated SOC–SOH Estimation Framework for Whole-Life-Cycle Lithium-Ion Batteries. *Energy* **2024**, *288*, 129801.
16. Gu, X.; Ieromonachou, P.; Zhou, L.; Tseng, M.-L. Developing Pricing Strategy to Optimise Total Profits in an Electric Vehicle Battery Closed Loop Supply Chain. *J Clean Prod* **2018**, *203*, 376–385.
17. Zhu, X.; Yu, L. Screening Contract Excitation Models Involving Closed-Loop Supply Chains under Asymmetric Information Games: A Case Study with New Energy Vehicle Power Battery. *Applied Sciences* **2019**, *9* (1), 146.

DOI: 10.36838/v8i2.50

18. Gu, H.; Liu, Z.; Qing, Q. Optimal Electric Vehicle Production Strategy under Subsidy and Battery Recycling. *Energy Policy* **2017**, *109*, 579–589.
19. Guan, Y.; Hou, Q. Dynamic Strategy of Power Battery Closed-Loop Supply Chain Considering Cascade Utilization. *IEEE Access* **2022**, *10*, 21486–21496.
20. Deng, Z.; Xu, L.; Liu, H.; Hu, X.; Duan, Z.; Xu, Y. Prognostics of Battery Capacity Based on Charging Data and Data-Driven Methods for on-Road Vehicles. *Appl Energy* **2023**, *339*, 120954.
21. He, H.; Zhang, J.; Wang, Y.; Jiang, B.; Huang, S.; Wang, C.; Zhang, Y.; Xiong, G.; Han, X.; Guo, D.; others. EVBattery: A Large-Scale Electric Vehicle Dataset for Battery Health and Capacity Estimation. *arXiv preprint arXiv:2201.12358* 2022.
22. Palacín, M. R.; de Guibert, A. Why Do Batteries Fail? *Science (1979)* **2016**, *351* (6273), 1253292. https://doi.org/10.1126/science.1253292.
23. Al-Alawi, M. K.; Cugley, J.; Hassanin, H. Techno-Economic Feasibility of Retired Electric-Vehicle Batteries Repurpose/Reuse in Second-Life Applications: A Systematic Review. *Energy and Climate Change* **2022**, *3*, 100086.
24. China Association of Automobile Manufacturers. Ownership. 2023. http://www.caam.org.cn/chn/7/cate_120/list_1.html.
25. Ouyang, M. EV Ownership Predictions. 2024. http://finance.people.com.cn/n1/2024/0228/c1004-40185301.html.
26. EVChina. Deep Analysis of the Recycling and Utilization of Power Batteries. 2022. http://www.evinchina.com/newsshow-1122.html.

■ Authors

Jeffrey Shen is a junior at West Island School in Hong Kong. He aims to integrate mathematical modeling and economic theory to analyze technological trends and develop data-driven policy suggestions.

DOI: 10.36838/v8i2.50

IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Chronic Hypoxia Induces Cardiomegaly and Increased Blood Flow Velocity during Chicken Heart Development

Eugene Fedutinov,[1] Molly Robinson,[2] Gray Franey,[3] Shonali Chakravarty[4]

1) Grover Cleveland High School, 3400 SE 26th Ave, Portland, OR, 97202, USA; fedutinovkid@gmail.com
2) McMinnville High School, 615 NE 15th St, McMinnville, OR, 97128, USA
3) Portland Community College, Rock Creek Campus, 17705 NW Springville Rd, Portland, OR, 97229, USA
4) Jesuit High School, 9000 SW Beaverton Hillsdale Hwy, Portland, OR, 97229, USA

ABSTRACT: This research examines the effects of hypoxia—low oxygen conditions—on embryonic heart development, specifically on heart size and blood flow velocities as markers of further defects. These early markers can be associated with later defects that come from these conditions, and explore the implications of hypoxia in all species' development. Fertilized chicken eggs (*Gallus gallus domesticus*) wrapped in clay and aluminum foil were used to model hypoxic embryo conditions. Each egg was incubated until approximately HH31 (day 7) or approximately HH35 (day 9). Then each egg was windowed, imaged using an ultrasound (Vevo 2100), dissected, weighed, and imaged under a microscope. We used the Doppler ultrasound feature of the Vevo 2100 system to determine flow velocity along the heart outflow tract. ImageJ was used to find the lengths and widths of the hearts from microscopic images by calibrating the size of the heart to the size of a known object, in our case, a 0.385 mm wide wire. Our results show that embryos that developed in hypoxic conditions had both larger hearts and faster cardiac blood flow velocities than control embryos, demonstrating that development in hypoxic conditions leads to abnormal development—an enlarged heart and faster blood flow— that perhaps can be projected onto humans.

KEYWORDS:Biomedical Engineering, Cardiovascular System, Hypoxia, Chicken Embryos, Blood Flow Velocity.

Introduction

Problem Statement and Research Aim:

In cases of chronic hypoxia during pregnancy, the developing cardiovascular system may adapt by developing heart defects. The developmental plasticity of the embryo allows it to compensate for the decreased oxygen levels, but this adaptation can lead to ventricular septal defects (VSD), atrial septal defects (ASD), or patent ductus arteriosus (PDA), which are heart malformations at birth. In this study, we focused on the effects of chronic hypoxia.

Congenital heart defects (CHD) are present in about 1% of births. VSDs are the most common CHD and account for 37% to 64% of cases of CHD or about 0.3% of births.[1] Moreover, VSDs have been studied as a cause of chronic hypoxia rather than a side effect. Except in cases of pulmonary arterial hypertension (PAH), the mortality rate for VSDs post-operation is about 1%.[2] Blood flow (hemodynamics) during pregnancy is key in understanding the development of CHDs, such as VSD, as well as the redirection of blood flow due to malformations. During pregnancy, the placenta forms as a temporary organ and the source of nutrients, oxygen, and waste filtration for the developing fetus. The fetus is attached to the placenta by the umbilical cord and depends on the mother to act as a sole source of resources.[3] Because of the position within the uterus, the fetus is unable to breathe in oxygen with its own circulation system. The dependence on the mother for oxygen can lead to a variety of complications in the case of placental, umbilical, and maternal conditions that may cause chronic hypoxia. Compression of the umbilical cord restricts oxygen to the fetus, as does any issue concerning the placenta, such as placental infarction, altered development, or poor placental function. In cases of maternal diseases such as preeclampsia, hemoglobinopathy, pulmonary hypertension, anemia, substance abuse, high altitude during gestation, and low blood pressure, chronic hypoxia is inflicted upon the fetus.[4] During human pregnancy, cardiomyocytes (the muscle cells of the heart) begin to contract 16 days into development.[5] Since cardiomyocytes continue to contract during development, pumping blood to the embryo, embryogenesis depends on the circulation system. In cases of chronic hypoxia, partially oxygenated blood is circulated, affecting developing tissues and forcing the fetus to compensate by redirecting blood flow and even developing cardiomyocyte hypertrophy.[6-9] These developmental changes may affect the individual later in life, through cases of congenital heart disease.

Current research on hypoxia (using chickens as a model organism) has focused heavily on the rate of development, long-term outcomes, cardiac mass and proportions, epigenetics, and resulting heart defects.[10-13] More recently, doctoral student Nina Kraus from the University of Vienna, Austria, established a procedure using clay to emulate chronic hypoxia in chicken embryos. Because chicken embryos breathe through their egg shell, the clay creates hypoxic conditions. The goal of our research is to identify how chronic hypoxia affects cardiac development, blood flow, and cardiac anomalies in chicken embryos. The focus on hemodynamics, as opposed to long-term outcomes and heart mass, has been stressed in this research.

DOI: 10.36838/v8i2.61

Chicken Embryos as Model Organisms:

Avian embryos are frequently used as models of cardiac development because they have a relatively quick incubation period of 21 days, and their hearts develop in a similar pattern to the human heart, resulting in a four-chamber configuration. Developmental programs, moreover, are highly conserved in vertebrate species. Chicken embryos can thus be used to study the effects of chronic hypoxia on hemodynamics and heart development. Hamburger-Hamilton (HH) Stages identify the progression of chicken development during incubation. The method of staging comes from a study done by Hamburger and Hamilton in 1951. Images of the stages are shown in Figure 1.

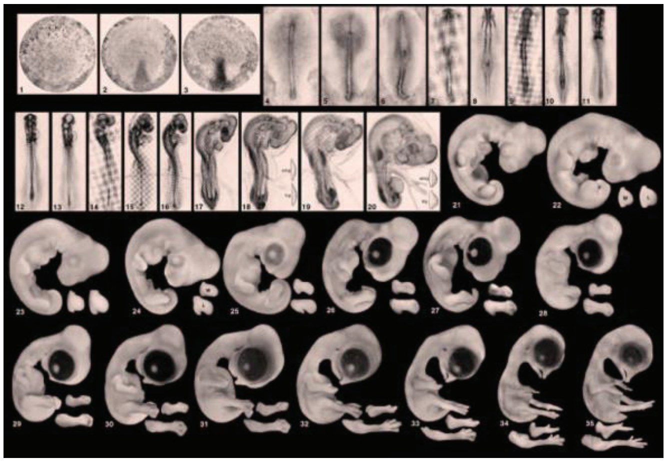


Figure 1: Hamburger Hamilton stages of chick embryo development. Used to accurately remeasure stages of growth throughout the experiment.[14]

HH stages identify how developing embryos change over time (during the 3 weeks of incubation) and the distinct characteristics they develop. There are 46 total stages of chick incubation, with HH46 being the final one and representing a hatched chick. At stage HH10 of embryo development, cardiac cells of a primitive tubular heart begin to contract, establishing embryonic circulation.[15] At HH31 (approximately day 7), the outflow tract continues to septate into vessels to provide oxygen and nutrients to the developing organs. The coronary vessels are evenly spaced, and the aortic and pulmonary valve leaflets have changed position to angle in towards each other. At HH35 (approximately day 9), the outflow tract and ventricles have fully septated, and the semilunar valves have finished developing. Cardiac neural crest cells spread through the cardiac plexus (the network of nerves located at the base of the heart) to prepare for the development of the peripheral conduction system (the Purkinje fibers). During embryological development, the chicken embryo receives oxygen through a system of gas exchange. When the egg is laid, the inner membrane shrinks slightly, creating an air pocket at the blunt end of the egg. This area grows larger as the egg ages, due to moisture diffusing out of the shell, and needs to remain uncovered for proper growth. Gas diffusion occurs through the pores on the shell of the egg, and therefore, claying (or covering) around half of these pores would create a hypoxic environment, without completely cutting off oxygen.[16] In our research, we used chicken embryos as our model organisms to study the effect of chronic hypoxia on cardiac development. To do so, we used clay wrapped around the bottom portion of the egg (blunt side up). Gas exchange occurs through the shell into the area between the inner and outer membranes, so the clay restricts access to oxygen (without completely blocking it), simulating chronic hypoxic conditions characterized by a reduction in oxygen for an extended period of development.

■ Methods

Fertilized chicken eggs were first clayed at day 2 or 3 of incubation to simulate hypoxic conditions. They were then incubated until day 7 or day 9, when they were imaged through an ultrasound to measure blood velocity through the heart. Once imaged, the embryos were dissected to remove the hearts, which were then either frozen in optimal cutting temperature compound (O.C.T. gel) in preparation for histological slicing to determine structure under a microscope or preserved in phosphate-buffered saline (PBS). The hearts that were preserved in PBS were then weighed. The frozen hearts were cut via cryostat in 10 µm thick slices, then stained using either a hematoxylin and eosin (H&E) stain or a Polysciences Differential Quik Stain. Under the microscope, they were measured in comparison to a 0.385 mm craft wire to find the length and width of each embryo's heart. We performed experiments in three sequential batches, or trials, adjusting techniques as we were learning to perform the studies and as needed. All eggs were incubated in an approximately 37°C incubator with 65%-80% humidity.

Claying:

Trial 1 included 26 eggs, which were split into 13 controls and 13 experimental clay-covered eggs. To begin trial 1, the bottom half of the experimental clay-covered eggs was wrapped with clay on day 3, then aluminum foil (to prevent flaking), leaving the blunt end uncovered. This created a hypoxic environment by eliminating oxygen diffusion through half of the shell. At day 3 of incubation, the embryos were at approximately HH18, where the heart, which is tubular at this stage, is in an S-shaped loop. The 13 control eggs were not covered with clay. All 26 eggs were windowed on day 3 using curved and straight forceps, and the outer embryo membrane was removed using forceps to make the embryos visible. The windows were then covered with plastic wrap and secured with glue, see Figure 2 (left). Trial one included 2 experimental groups: control and day 3 clayed embryos (clayed D3).

Trial 2 included 21 eggs, which were split into 5 controls and 16 experimental clay-covered eggs. Of the 16 clay-covered eggs, 8 were covered with clay on day 2 of development, and 8 were covered with clay on day 3 of development (Figure 2). At day 2, eggs were at approximately HH stages 12-13, where dextral looping of the heart begins. At these stages, the endocardial cushions also emerge (precursors to valves). The trial 2 eggs were kept unwindowed until day 7 of incubation, see Figure 2 (right). At day 7, the eggs were at stage HH31-32. By HH31, the distal portion of the outflow tract has finished septating, the coronary arteries and veins are in their final positions, and the aortic and pulmonary valves are angled.[15] Trial 2 included 3 experimental groups: control, day 2 clayed embryos, and day 3 clayed embryos.

Trial 3 included 15 eggs, which were split into 6 controls and 9 experimental clay-covered eggs. All 9 experimental eggs were covered with clay on day 3 of development. The trial 3 eggs were kept unwindowed until day 9 of development. On day 9 of development, the eggs were around stage HH35-36, at which the Phalanges in the toes developed.

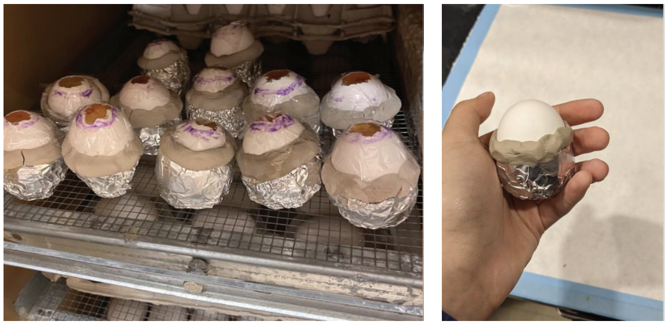


Figure 2: Trial 1 eggs (left) are shown windowed and clayed. Trial 2 eggs (right) are shown only clayed. Used as a model for what a recreatable setup would include.

Experimental groups:

In 7 groups were used in this experiment: i) control incubated until day 7 (windowed), ii) clayed day 3 and incubated until day 7 (windowed), iii) control incubated until day 7, iv) clayed day 2 and incubated until day 7, v) clayed day 3 and incubated until day 7, vi) control incubated until day 9 and vii) clayed day 3 and incubated until day 9. Trial 1 included control incubated until day 7 (windowed) and clayed day 3 and incubated until day 7 (windowed). Trial 2 included control incubated until day 7, clayed day 2, and incubated until day 7, and clayed day 3 and incubated until day 7. Trial 3 included a control incubated until day 9 and clayed on day 3 and incubated until day 9.

Ultrasound:

On day 7 or 9, eggs were taken out of the incubator to perform in vivo ultrasound imaging via a FujiFilm VisualSonics Vevo 2100 Imaging System. Trial 1 eggs, which were previously windowed, were windowed further to increase space for the ultrasound transducer. In trials 2 and 3, eggs were windowed for the first time. The windowed egg was then placed into a special 3D printed egg holder atop a rising platform, which had the transducer mounted above. The egg was then raised so that the embryo could touch the transducer. Using the visual displayed on the ultrasound screen (Figure 3), the egg was positioned by hand to get the right angle. To measure the pulse wave velocity, the heart outflow tract, which directs blood from the heart to the circulation, had to be clearly visible, and the color Doppler feature of the Vevo system helped find it (Figure 4). Once the outflow tract was identified, the ultrasound's Vevo Lab software was used to measure blood flow velocity through the outflow tract. Peak and mean velocities (mm/s) were taken from three pulses (Figure 5). This process was repeated for each egg. Some eggs could not be imaged properly due to their position or death during the process.

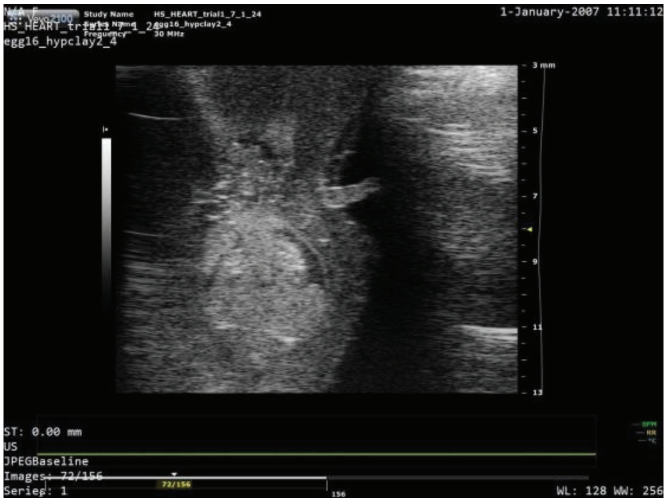


Figure 3: Normal B-mode scan with a 4-chamber view of the heart. Shows a regular-sized heart and is used as a model to ensure that all scans are imaged at similar angles to get reproducible results.

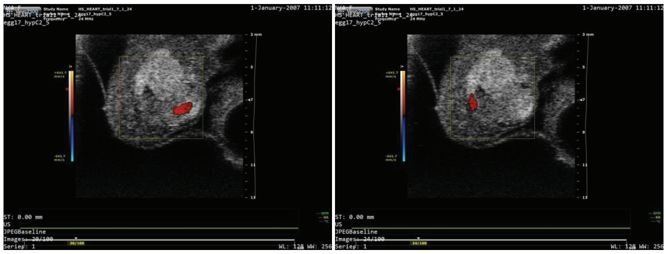


Figure 4: Colored Doppler images. Red indicates blood flow toward the transducer. Blue indicates blood flow away from the transducer. Shows imaging angle for reproducible results. Shows which heart valves were measured to achieve velocity measurements.

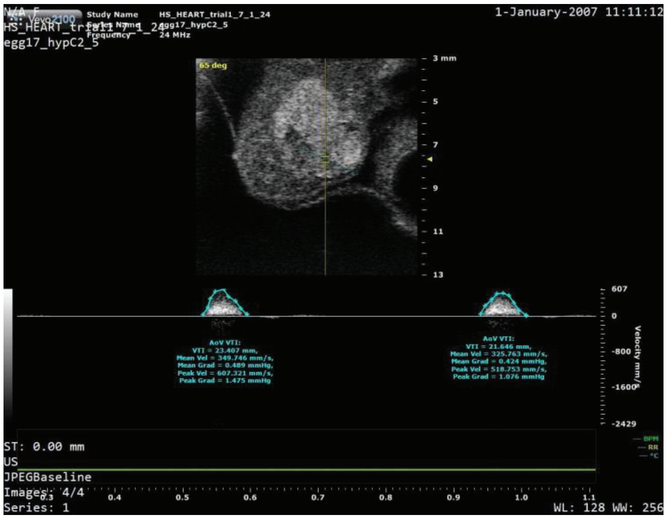


Figure 5: Pulse wave doppler mode, velocity measurement. Shows how average and peak velocities were obtained for reproducibility. Shows the heart of a hypoxic embryo with relatively high blood flow velocity.

Dissection:

Once an embryo's in vivo ultrasound imaging was complete, the embryo was removed from its shell for dissection. The embryo was transferred from its shell to a petri dish (Figure 6) using a scoopula and then decapitated using a scalpel. The ribcage was then cut open using microdissection scissors and peeled away using forceps, exposing the heart (Figure 7).

DOI: 10.36838/v8i2.61

Scissors were used to cut around the heart, removing as much excess tissue as possible. Once completely separate, a transfer pipette was used to place 1-3 drops of potassium chloride (KCl) on the heart to ensure it would stop beating and the muscle would fully relax. Finally, the heart was placed in a Tissue-Tek cryomold and covered with O.C.T. gel before freezing. Samples were frozen via dry ice and then placed in a -80°C freezer.

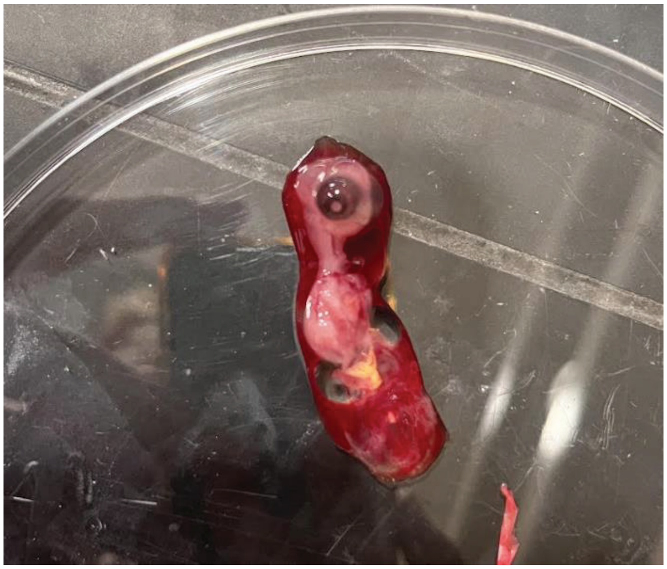


Figure 6: Chicken embryo removed from egg before dissection. The embryo is on day 9 of incubation. Shows how chicken embryos were analyzed for abnormalities that Nina Kraus found.

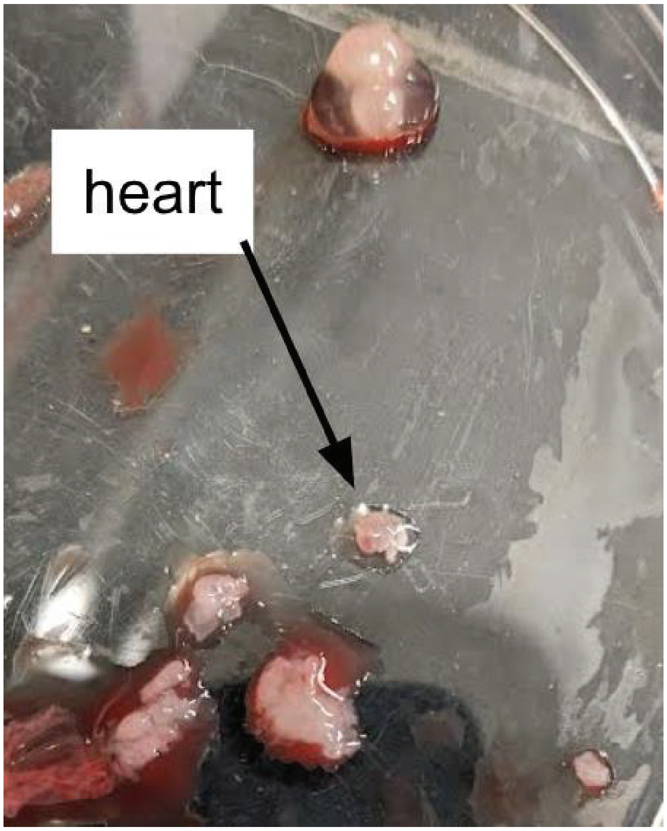


Figure 7: Labelled heart post-dissection. The heart is from day 9 of incubation. Shows how a hypoxic heart was dissected for reproducibility.

Cryostat:

Once frozen, a Leica CM1860 cryostat (Figure 8) was used to cut the O.C.T. block with the frozen hearts embedded in it into 10 µm thick slices for histology slide preparation. First, the cryostat's slice thickness was set to 50 µm. The heart, covered in a frozen square of O.C.T. gel, was taken out of its cryomold and placed into the cryostat's mount. The wheel on the side of the cryostat was cranked to move the mount down into the blade, to cut away excess O.C.T. gel. Once the desired amount was removed, the glass anti-roll shield and blade were fixed in place, and the slice thickness was set to 10 µm. Once one slice was cut, it was placed on a glass slide, and the O.C.T. was allowed to melt while the tissue stuck to the glass. The process was repeated until the whole heart was sectioned.

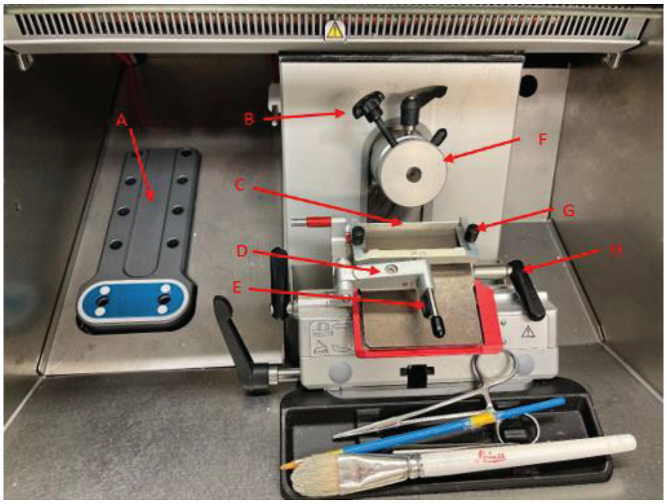


Figure 8: View of Leica CM1860 cryostat work area. (A) Tissue storage area. (B) Tightening knob. (C) Microtome blade. (D) Glass anti-roll plate holder. (E) Anti-roll plate adjustment knob. (F) Chuck. (G) Glass plate tightening screws. (H) Microtome blade clamp lever. Used for reproducibility.

Staining:

There were two different methods of staining that we used, and heart sections were stained with one or the other (it is not possible to stain with both simultaneously). The first was a basic hematoxylin and eosin (H&E) histological stain. The H&E staining process included placing the heart slides in baths of different chemicals and letting them sit for different time intervals. This included two baths of phosphate-buffered saline (10 minutes each), one of hematoxylin (3 minutes), a rinse of deionized water, then one bath of eosin (30 seconds), one of 50% ethanol (1 minute), one of 90% ethanol (1 minute), two of 100% ethanol (1 minute each), and two of xylene (10 minutes each). Once complete, the stain highlighted cell nuclei, cytoplasm, and extracellular matrix. Slides were then covered with cover slips and were ready for imaging.

The second staining method used the Polysciences Differential Quik Stain, which had only three steps. The heart slides were first dipped into the fixative (methanol), then Solution A (an eosin-based stain), then Solution B (methylene blue). The slides were then covered with cover slips and ready for imaging. This stain highlighted nearly all cells.

Microscope:

The microscope used in this experiment was a Leica stereo microscope M125 C, along with the pco.camware that recorded movies or still images of the hearts. ImageJ (an open-source, free imaging software) was used to measure the length and width of each embryo's heart from the cryo-sections. With the slide positioned under the microscope, a piece of 0.385 mm silver-plated craft wire was placed on the slide next to a chosen heart. The camera was zoomed in to view the chosen heart and wire, and the focus was adjusted to view both objects clearly. An image was captured and opened in ImageJ, where lines were drawn across the wire diameter, length of heart, and width of heart (Figure 9). Lines were measured (in pixels) of each of the lines, and a ratio was created of the wire diameter in pixels and millimeters. The ratio was then used to calculate the heart width and length in millimeters.

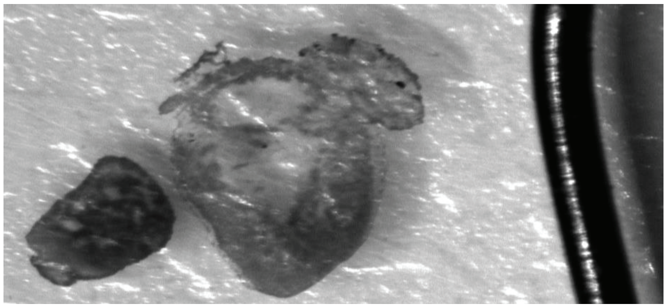


Figure 9: Image of day 7 embryo heart from pco.camware software. 0.385 mm wire located on the right side. Shows a relatively large hypoxic heart and the process for analysis for reproducibility.

Weighing:

6 hearts incubated until day 7(1 control, 3 clayed day 2, 2 clayed day 3) and 4 hearts incubated until day 9 (1 control, 3 clayed day 3) were preserved in tubes with PBS. To weigh, a Sartorius Precision Scale was zeroed to the weight of one weigh boat. Then, a heart was removed from its tube using straight forceps and placed onto the weigh boat on the scale. Mass was recorded in grams. The process was repeated for each heart.

Result and Discussion

Because eggs were windowed early (day 3) for trial one, and we suspect the window enabled additional oxygenation through the window, we only used the results of trials 2 and 3 for analysis. Instead, we used trial 1 embryos for practice and to test our techniques.

Our data show that chicks inside eggs wrapped in clay to simulate hypoxia had (at day 7) a faster blood flow velocity through the outflow tract than control embryos, both for embryos inside eggs clayed at day 2 or day 3. Eggs were divided into 3 experimental groups: control, clayed on day 2, and clayed on day 3. On day 7 of incubation, the average mean blood flow velocity increased by 62% from the control embryos for embryos clayed on day 2 of incubation and increased by 66% from the control embryos for embryos clayed on day 3 of incubation. Figures 10 and 11 show graphs of average mean velocities and average peak velocities to display this behavior. These results show that hypoxic embryos had, on average, developed both a faster mean blood flow and a higher peak blood flow when compared to the control embryos. But there was no significant difference between eggs clayed on day 2 and eggs clayed on day 3. Figure 10 has outliers, which were better addressed in Figure 11 by averaging out the results.

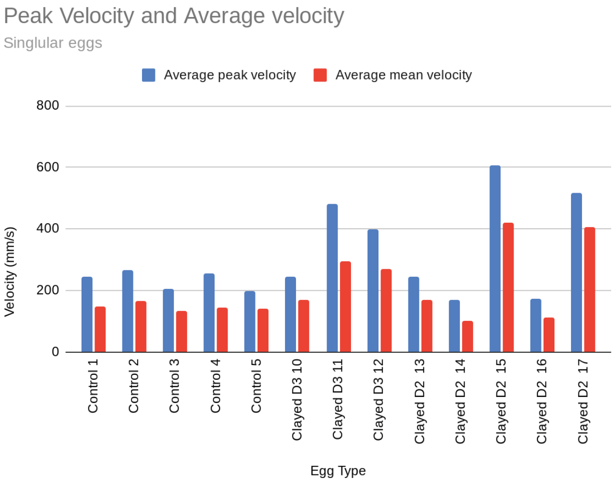


Figure 10: Individual measurements of peak and mean blood flow velocities through the ventricular outflow tract at day 7. Higher peak and average blood flow velocities can be seen among the embryos that were clayed.

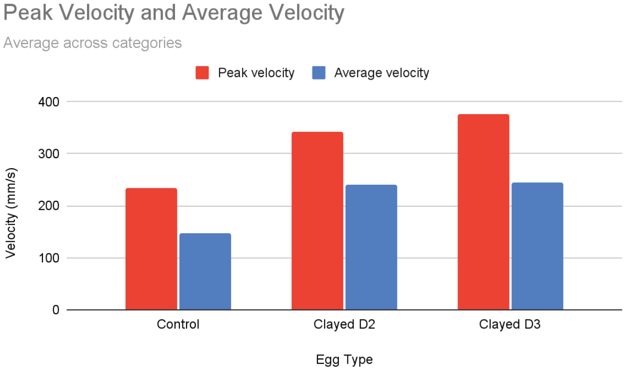


Figure 11: Measured blood flow velocities of the outflow tract of day 7 chick embryos. When averaged across categories, there is a significant increase in velocity among clayed eggs.

We also measured the same data for eggs that incubated until day 9 of development and found similar results. In Figures 12 and 13, we can see that both the mean and peak velocities stayed higher than the control embryos. This data shows that the higher blood velocity did not go away with development. Higher mean and peak velocities persisted over time for hypoxic embryos.

DOI: 10.36838/v8i2.61

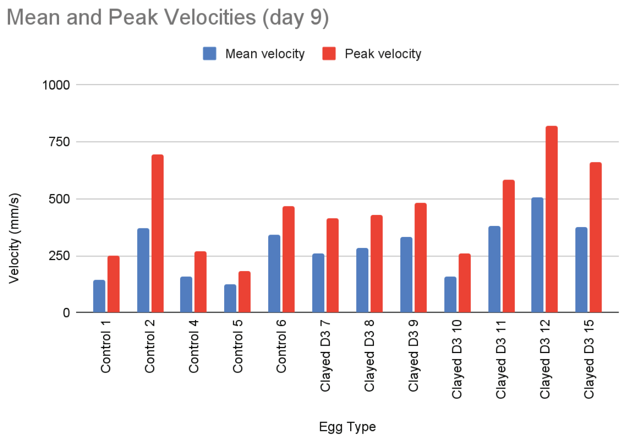


Figure 12: Individual measurements of blood velocities through ventricular outflow (day 9). Among the eggs incubated until day 9, there is an increased blood flow velocity among the clayed eggs, with some outliers in the control group.

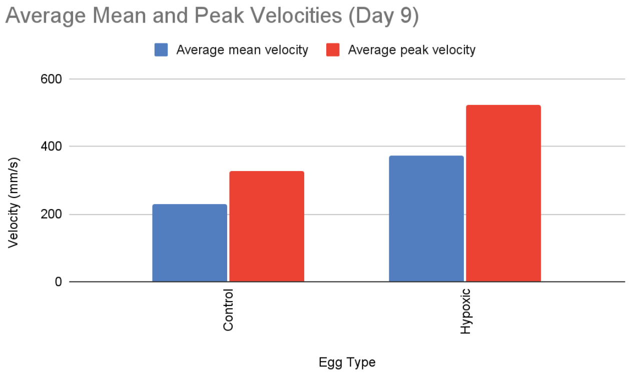


Figure 13: Average blood flow velocities of the outflow tract of day 9 chick embryos. When averaged across groups, there is a significant increase in the blood flow velocities among the clayed/hypoxic eggs.

Additionally, we measured the length and width of the clayed and control embryos' hearts. From stained sections of the hearts, approximately corresponding to the middle of the heart, where the heart is larger and wider, we found that, on average, the hypoxic embryos developed both longer and wider hearts. Figures 14 and 15 show that the hypoxic (clayed) embryo hearts were larger than the control embryo's hearts. While cutting the hearts to put onto the microscope slides, some of the hearts were sectioned in an incorrect orientation, and therefore, we could not get both length and width, but only one or the other. That is why some heart measurements only have length or width measurements. Also, some embryos were not processed for staining, hence we have data from fewer embryos than for previous measurements. These results show that the hypoxic embryos' hearts developed to a larger size. There was not enough data to determine differences between eggs that were clayed on day 2 and eggs that were clayed on day 3 of development.

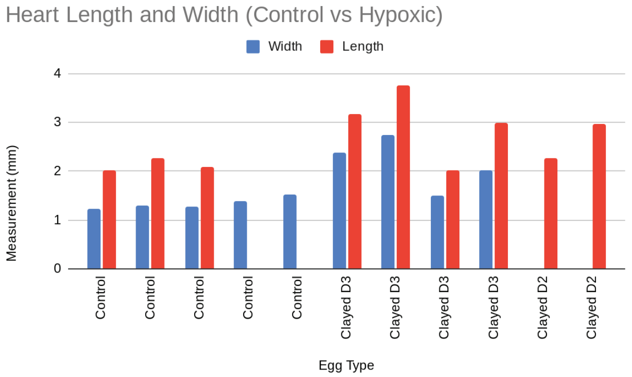


Figure 14: Individual measurements for heart sizes for each group of eggs (day 7). It can be seen that there was a slight increase in the heart size among the eggs that were clayed compared to the control group.

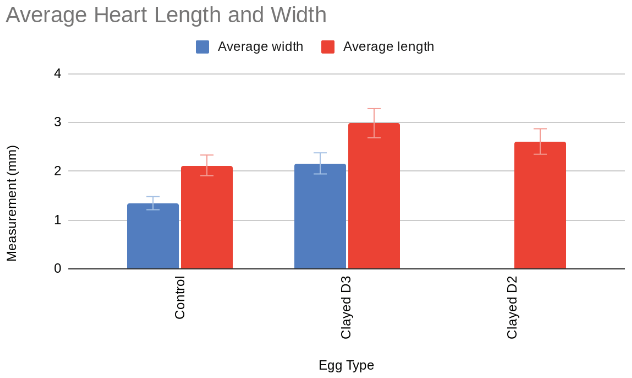


Figure 15: Average heart size for each group of embryos (day 7). When averaged across groups, a slight increase can be seen in the heart size of the control/hypoxic eggs when compared to the control groups.

Our results also show that the hypoxic embryos developed slightly heavier hearts. Not all the hearts that were dissected were weighed, so the data set is smaller than the other experiments, and only includes 1 control heart, which is not ideal. Figures 16 and 17 show that, on average, the hearts that developed in hypoxic conditions weighed more than those that developed normally. This shows the larger heart size was due to an actual increase of cardiac mass in the hypoxic embryos. These findings are significantly hindered by a limited sample size.

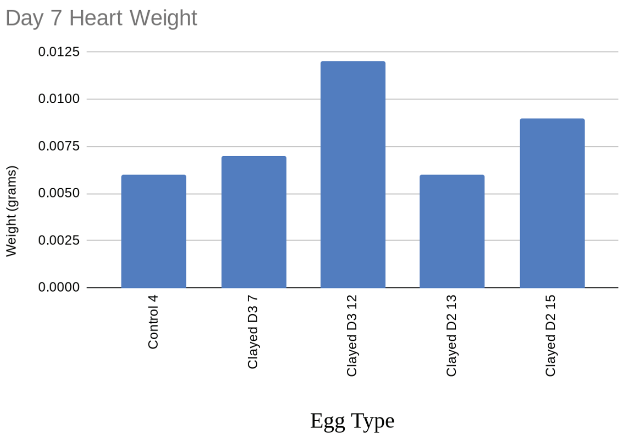


Figure 16: Individual measurements of the weight of the heart for each group (dissected on day 7). This figure shows a slight increase in the cardiac mass among the clayed embryos when compared to the control groups. This figure supports the previous findings and implies that the increase in cardiac size is due to increased cardiac mass.

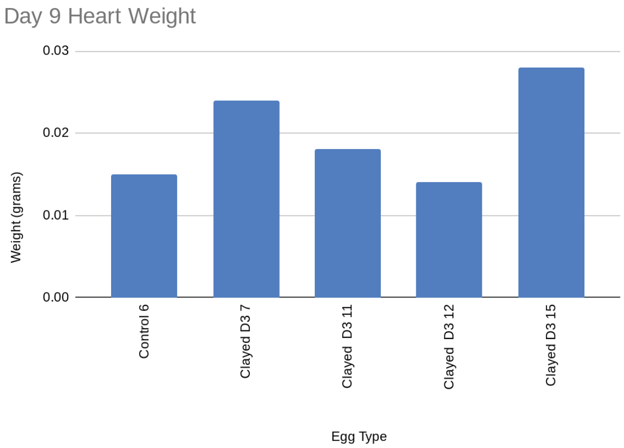


Figure 17: Individual weights of hearts for each group (dissected day 9). Similar to the previous figure, this shows a slight increase in cardiac mass among the clayed eggs and implies that the increase in heart size is due to increased cardiac mass.

The first trial of windowed eggs could not be used for data, as the method for windowing the eggs exposed the embryos to excess air, which made the trial unsuitable to induce hypoxia or to provide a suitable control reference. The sample size in this experiment was limited due to several issues, such as i) one of the incubators that was used did not maintain ideal humidity due to frequent use (opening of the door) and decreased viability, ii) we could not measure the ratios of heart size to embryo size due to no measurements of embryos. This research was also done on a limited 3-week period, which did not allow for as much testing as we wanted, and limited time to incubate eggs. In further research, the same experiment could be conducted with a larger sample size (15-30 viable chicken embryos per experimental group), and by leaving the eggs to incubate longer (until at least day 12/13), observations of further heart defects could be conducted.

Using Welch's t-test to test for statistical significance, it was found that when comparing control vs clayed egg blood flow velocities, both peak and average velocities attained a value of $p<0.01$ ($p=0.0022$ and $p=0.001$, respectively, along with t values of –3.39 and –3.79). This signifies that the results are significant, and both measures of blood flow velocities were significantly higher in clayed groups. For weight, there were not enough samples to run a statistical significance test, so more samples would need to be taken.

■ Conclusion

Mothers with placental issues or locational issues that can cause hypoxia are more likely to deliver babies with congenital heart defects. Around 1% of babies in the US are born with congenital heart defects, and 1 in 4 of these are critical heart defects.[16] These congenital heart defects form in the early stages of embryonic development and are very susceptible to different variables, especially a mother's placental health and choices, which can affect both changes in the embryo's access to oxygen and blood flow in the embryo's heart. Both reduced oxygen and altered blood flow in hypoxic embryos can lead to heart defects. Previous research has not investigated variability in heart sizes and weights in early stages of life due to hypoxic conditions, nor changes in blood flow velocities in the heart. To our surprise, we found that hypoxic hearts are larger than control hearts (cardiomegaly), and blood flow velocities in hypoxic embryos are increased with respect to the control. We did not notice other malformations with reference to control hearts.

In our research, we used chicken embryos to model heart development and study the effect of hypoxia on heart development. Based on our experimental results, we found that embryos that develop in hypoxic conditions have altered heart development, but more measurements will need to be done in the future to confirm our results. Along with this, this experiment is limited due to the use of wet measurements for heart weight; future experiments could include dry measurements to determine further if heart growth was due to edema or tissue growth. This experiment could also be expanded to other animal models—quail, zebrafish, mice—to observe if similar results occur. More measurements could be taken to study other complications. Running this experiment several times could eliminate some errors, increase our certainty, and provide better insight into hypoxic heart development.

■ Acknowledgments

The authors would like to thank the Rugonyi Lab at the Oregon Health and Science University (OHSU) and the Kheradvar Lab at the University of California-Irvine (UCI). We are also grateful to Nina Kraus, who shared with us the egg-claying method to induce hypoxia. This research was conducted as a part of the HEART (High School Students Engaged in Applied Sciences, Research, and Technology) internship program at OHSU, funded by the National Science Foundation (NSF) and the OHSU quantitative and systems biology (QSB) program. The authors would like to acknowledge Dr. Sandra Rugonyi, Scott Howsden, and Makena Phillips for their mentorship and assistance in conducting this research.

References

1. Villines, Z.; Stephens, C. *What is a ventricular septal defect (VSD)*? www.medicalnewstoday.com. https://www.medicalnewstoday.com/articles/vsd-heart#causes (accessed 2024-07-10).
2. Dakkak, W.; Oliver, T. I.; Alahmadi, M. *Ventricular Septal Defect*; StatPearls Publishing, 2023.
3. Cleveland Clinic. *Placenta: Overview, Anatomy, Function & Complications*. Cleveland Clinic. https://my.clevelandclinic.org/health/body/22337-placenta.
4. Silvestro, S.; Calcaterra, V.; Pelizzo, G.; Bramanti, P.; Mazzon, E. Prenatal Hypoxia and Placental Oxidative Stress: Insights from Animal Models to Clinical Evidences. *Antioxidants* **2020**, *9* (5), 414. https://doi.org/10.3390/antiox9050414.
5. Tyser, R. C.; Miranda, A. M.; Chen, C.; Davidson, S. M.; Srinivas, S.; Riley, P. R. Calcium Handling Precedes Cardiac Differentiation to Initiate the First Heartbeat. *eLife* **2016**, *5*. https://doi.org/10.7554/elife.17113.
6. Fan, C.; Iacobas, D. A.; Zhou, D.; Chen, Q.; Lai, J. K.; Gavrialov, O.; Haddad, G. G. Gene Expression and Phenotypic Characterization of Mouse Heart after Chronic Constant or Intermittent Hypoxia. *Physiological Genomics* **2005**, *22* (3), 292–307. https://doi.org/10.1152/physiolgenomics.00217.2004.
7. Bae, S.; Xiao, Y.; Li, G.; Casiano, C. A.; Zhang, L. Effect of Maternal Chronic Hypoxic Exposure during Gestation on Apoptosis in Fetal Rat Heart. *American Journal of Physiology-Heart and Circulatory Physiology* **2003**, *285* (3), H983–H990. https://doi.org/10.1152/ajpheart.00005.2003.
8. Martin, C.; Yu, A. Y.; Jiang, B.-H.; Davis, L.; Kimberly, D.; Hohimer, A. Roger.; Semenza, G. L. Cardiac Hypertrophy in Chronically Anemic Fetal Sheep: Increased Vascularization Is Associated with Increased Myocardial Expression of Vascular Endothelial Growth Factor and Hypoxia-Inducible Factor 1. *American Journal of Obstetrics and Gynecology* **1998**, *178* (3), 527–534. https://doi.org/10.1016/s0002-9378(98)70433-8.
9. Val'kovich, E. I.; Molchanova, V. V.; Davydova, M. K.; Davydova, O. K. [Changes in the Myocardium of Fetuses and Newborn Infants as a Result of Hypoxia]. *Arkhiv Anatomii, Gistologii I Embriologii* **1986**, *90* (3), 35–39.
10. Haron, A.; Ruzal, M.; Shinder, D.; Druyan, S. Hypoxia during Incubation and Its Effects on Broiler's Embryonic Development. *Poultry Science* **2021**, *100* (3), 100951. https://doi.org/10.1016/j.psj.2020.12.048.
11. Wikenheiser, J.; Wolfram, J. A.; Gargesha, M.; Yang, K.; Karunamuni, G.; Wilson, D. L.; Semenza, G. L.; Agani, F.; Fisher, S. A.; Ward, N.; Watanabe, M. Altered Hypoxia-Inducible Factor-1 Alpha Expression Levels Correlate with Coronary Vessel Anomalies. *Developmental dynamics: an official publication of the American Association of Anatomists* **2009**, *238* (10), 2688–2700. https://doi.org/10.1002/dvdy.22089.
12. Wikenheiser, J.; Doughman, Y.-Q.; Fisher, S. A.; Watanabe, M. Differential Levels of Tissue Hypoxia in the Developing Chicken Heart. *Developmental Dynamics* **2005**, *235* (1), 115–123. https://doi.org/10.1002/dvdy.20499.
13. Su, Z.; Liu, Y.; Zhang, H. Adaptive Cardiac Metabolism under Chronic Hypoxia: Mechanism and Clinical Implications. *Frontiers in Cell and Developmental Biology* **2021**, *9*. https://doi.org/10.3389/fcell.2021.625524.
14. Hamburger, V.; Hamilton, H. L. A Series of Normal Stages in the Development of the Chick Embryo. *Journal of Morphology* **1951**, *88* (1), 49–92. https://doi.org/10.1002/jmor.1050880104.
15. Martinsen, B. J. Reference Guide to the Stages of Chick Heart Embryology. *Developmental Dynamics* **2005**, *233* (4), 1217–1237. https://doi.org/10.1002/dvdy.20468.
16. Divya, D.; Khan, M.; Hussaini, J.; Ashgar, M.; Raushan, S. How Does the Chick Breathe inside the Shell? *International Journal of Scientific Research in Science, Engineering and Technology* **2019**, *1* (6).
17. CDC. *Data and Statistics*. Congenital Heart Defects (CHDs). https://www.cdc.gov/heart-defects/data/index.html.

Authors

Eugene Fedutinov is a junior at Cleveland High School. He is interested in the sciences, especially biology and organic chemistry. In his undergraduate studies, he hopes to conduct more research, become part of a lab, and eventually have a profession in a medical setting to help better the health of everyone.

Molly Robinson is a senior at McMinnville High School. She is interested in biomedical engineering, chemistry, and public health. As an undergraduate student, she hopes to study Public Health and one day pursue a medical degree to work in pediatrics.

Gray Franey is a senior at Early College High School located at Portland Community College. They enjoy biology, mathematics, and environmental conservation. In their undergraduate studies, they would like to study genetics and microbiology, with the goal of biomedical research.

Shonali Chakravarty is a senior at Jesuit High School. She has a passion for biology, psychology, and neuroscience. During her undergraduate studies, she hopes to participate in research and intern at hospitals.

IJHSR ijhighschoolresearch.org

RESEARCH ARTICLE

Food Preservation: Use of Silk Fibroin as an Edible Coating on Bananas

Matthew B. Martinelli

Highland Park High School, 4220 Emerson Ave., Dallas, TX 75205; matthewbmartinelli@gmail.com

ABSTRACT: Silk fibroin, the main structural protein of silk, is a biomaterial that has been extensively studied in textile, biomedical, and electronic applications. Due to its hydrophobic nature and its ability to assemble into antiparallel, beta-pleated sheets, the strength, flexibility, and conformability of fibroin have also led to its use as a coating to prolong produce. This study investigates whether the application of a 1% weight/volume aqueous silk fibroin suspension to bananas would prolong their shelf life; additionally, it attempts to determine whether increasing the beta-sheet content of the protein via water annealing (exposure of fibroin to water vapor in a vacuum at constant temperature) would be more effective than coating alone. Measured parameters of fruit spoilage assessed daily in the study were: visible signs of ripening, weight loss, maintenance of turgor, and mold/yeast growth. After the 8-day ripening period, bananas coated in fibroin and stored ambiently at room temperature exhibited less visible signs of spoilage, lost less weight, maintained more firmness/turgor, and grew fewer molds and yeast than did the control cohort. The effect of all measured parameters was slightly more pronounced in the group of bananas that were coated and annealed. Considering that one-third of the food produced in this country is never eaten, this research holds the promise of safely and effectively enhancing food preservation methods, with the hope of far-reaching societal impacts.

KEYWORDS: Biochemistry, Structural Biochemistry, Food Preservation, Silk Fibroin, Water-Annealing, Beta-Pleated Sheet.

Introduction

According to a recent study published by the US Environmental Protection Agency,[1] more than one-third of the food produced in this country is never eaten, wasting the resources used to produce the food and creating a variety of downstream environmental impacts. Additionally, regarding fruits and vegetables specifically, the Food and Agriculture Organization of the United Nations estimates a fifty percent loss of crops throughout the food supply chain, generally concentrated in the post-harvest, distribution, and end-user consumption stages.[2] Food waste contributes to food insecurity, reduced economic efficiency, and impaired efforts to address energy conservation and climate change. Improving the post-harvest shelf life of foods will not only lessen the need for new food production, but also reduce projected deforestation, biodiversity loss, greenhouse gas emissions, and water pollution and scarcity.

Many treatments to extend the shelf life of perishable produce have been explored, including cryopreservation, treatment with synthetic fungicides, modified atmospheric packaging, and osmotic and temperature treatments. Edible coating materials have also shown utility in food preservation, specifically those that demonstrate biocompatibility, biodegradability, antibacterial and antifungal properties, membrane-forming ability, and safety. Currently utilized coating materials include proteins, polysaccharides, resins, and lipids; however, each lacks some of the requisite physical and chemical characteristics of the ideal fruit and vegetable coating to prolong shelf life. Some desirable physical characteristics of an edible preservative coating include: barrier properties to control gas and moisture transfer; mechanical and tensile strength to resist cracking; and transparency to maintain the natural appearance of the fruit. Chemically, an ideal edible coating should have antimicrobial and/or antioxidant properties to impair or slow the growth of pathogens responsible for fruit spoilage, and should preferably be naturally occurring materials.

Silk fibroin is a biomaterial that has been extensively studied for its potential in textile, biomedical, and electronic applications.[2-4] It has been affirmed by the U.S. Food and Drug Administration (FDA) as a non-toxic substance and is classified as generally recognized as safe (GRAS) for its intended use as a surface-finishing agent on food. Silk fibroin is the main structural component of silk and is produced by the silkworm, *Bombyx mori (B. mori)*. The properties of silk fibroin are derived from its structure, which consists of hydrophobic blocks separated by hydrophilic acidic spacers.[2,4,5] In its natural state, silk fibroin forms antiparallel beta-pleated sheets as its secondary structure, providing strength and resilience to the protein. Additionally, the beta-pleated sheet structure of fibroin, combined with its inter- and intra-molecular hydrogen bonding, confers high flexibility and conformability of the protein. Notably, prior studies have shown that the beta-pleated sheet content of fibroin can be regulated and adjusted based on the time and temperature at which fibroin is exposed to water vapor or other polar solvents in vacuum conditions (annealing).[2,6]

Based on prior studies using silk fibroin as an edible coating for perishable food preservation, and considering that a 1% weight/volume fibroin solution was noted to be safe and efficacious in those studies, we sought to utilize the intrinsic properties of silk fibroin, such as its hydrophobicity and conformability, to design a 1% weight/volume water-based suspension of fibroin that assembles on the surface of miniature bananas upon coating. We also increased the beta-pleated

DOI: 10.36838/v8i2.69

sheet content of the fibroin via water annealing to determine whether increasing the content of beta-pleated sheets would confer additional benefit in prolonging the spoilage of the bananas. Six roughly identical miniature bananas were used in the study: two without fibroin added (control), two with fibroin solution coating, and two coated with the same concentration of fibroin and then water annealed for twelve hours. The bananas' physical appearance, weight loss, loss of turgidity, and presence of microbial growth were then recorded for all bananas as indicators of spoilage. This study is designed to contribute to the growing body of literature on safe, effective, and readily available preservative coatings to help address the critical issue of food spoilage.

■ Methods

Sample and Coating Preparation:

Six non-ripe miniature bananas were purchased, ensuring they were roughly the same sizes, shapes, and textures, and had similar degrees of mechanical damage. Similarities in color/ripeness, visible evidence of bruising, and texture/firmness to light pressure were subjectively assessed by the author. All bananas were weighed and photographed pre-experiment. Using a graduated cylinder, 1 g of silk fibroin powder (Advanced Biomatrix CytoSilk lypophilized silk fibroin) was dissolved in 100 mL of distilled water to generate a 1% weight/volume solution of silk fibroin in a clear plastic container. This was accomplished via gentle agitation and mixing until a homogeneous, clear, straw-colored solution was obtained. The bananas were labeled as follows: 2 bananas served as controls and were not treated with the fibroin or annealed ("Control #1" and "Control #2"); 2 bananas were coated with fibroin solution ("Dipped #1" and "Dipped #2"); 2 bananas were coated with fibroin solution, and then water annealed ("Dipped & Annealed #1" and "Dipped & Annealed #2").

All 4 were coated equally and uniformly with the fibroin solution by separately submerging each banana in the fibroin solution in a gallon-size Ziploc bag. Each banana was completely submerged in the solution for 60 seconds. All bananas were then placed on a drying rack at room temperature (approximately 21° C) for 4 hours.

Annealing:

After the 4-hour drying period, the 2 bananas labeled "Dipped & Annealed #1" and "Dipped & Annealed #2" were water annealed as follows. 100 mL of distilled water was poured into a vacuum-sealed bag. A small plastic container was inverted and placed at the bottom of the bag. The banana labeled "Dipped & Annealed #1" was placed on top of the container, with care being taken for the water not to contact the banana. The bag was then vacuum sealed using a food vacuum sealer. This process was then repeated for the banana labeled "Dipped & Annealed #2." Both vacuum-sealed bags were allowed to remain at room temperature (21° C) for 12 hours.

Storage and Sampling Cadence:

All bananas were stored at room temperature (21 °C) for the duration of the study period. Weights, in grams (taken on a digital scale) and digital photographs of each banana were obtained once daily for the duration of the study.

Weight Loss:

Weight loss of each banana, in grams, was calculated by subtracting the weight of each banana on day 8 of the study period from the initial weight of each. This data was recorded.

Turgidity and Resistance to Depression:

A 200 g weight was placed on each banana on day 8, and the degree of vertical depression by the weight, measured in millimeters, was photographed, measured, and recorded for each.

Fungal Plating and Enumeration:

On day 8, each of the bananas was swabbed, uniformly on all sides for 10 seconds, with sterile cotton-tipped applicators. Each sterile applicator was then used in a "back and forth" manner for a total of 5 seconds to plate material on labeled dichloran rose bengal chloramphenicol (DRBC) agar petri dishes, with care being taken to uniformly cover the surface of each plate. Microbial colony growth was manually counted and recorded for each petri dish after the 8-day period.

■ Results and Discussion

Results:

The four bananas treated with the 1% aqueous fibroin solution showed less visible signs of ripening and spoilage (i.e., less browning and bruising) than the two non-treated controls after the 8-day period. The two dipped and water annealed bananas also showed less visible signs of ripening than did their dipped, but non-water annealed counterparts **(Figure 1)**.

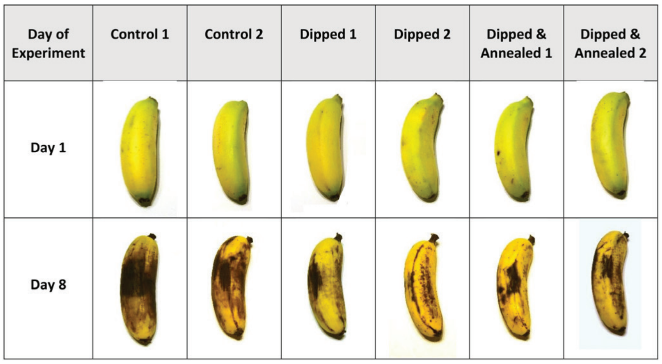


Figure 1: Digital photographs document the appearance of all 6 bananas between day 1 and day 8: 2 untreated controls, 2 dipped in 1% silk fibroin solution, and 2 dipped in 1% silk fibroin solution and subsequently water annealed. The most prominent visible signs of ripening occurred in the control group. The bananas that were dipped in silk fibroin and those that were dipped and water annealed maintained visible evidence of freshness longer, with less darkening, bruising, and shriveling seen.

Consistent with the volume loss that occurs during fruit ripening, all 6 bananas underwent some degree of weight loss during the study period. However, the most significant weight loss was exhibited in the untreated control cohort. Control bananas 1 and 2 showed a loss of 9.15 g and 7.34 g (or a 23.91% and 22.87% reduction in starting weight), respectively. Bananas 1 and 2 dipped in the silk fibroin solution lost 6.84 g and 7.26 g (18.86% and 19.66%), respectively. Bananas 1 and 2 that were dipped and then water annealed lost the least amount of weight at 6.54 g and 6.79 g (20.06% and 18.12%), respectively **(Table 1, Table 2, Figure 2, Figure 3)**. The average weight loss at the end of the study was 23.39% for controls, 19.26% for the dipped group, and 19.09% for the dipped and annealed cohort **(Table 3, Figure 4)**.

Table 1: Daily measured weights of all bananas on days 1 to 8 are shown. Calculated changes in weight between day 1 and day 8 are also displayed. Weight loss is a physical change associated with fruit ripening and spoilage. Consistent with the preservation effect of silk fibroin, those bananas dipped in silk fibroin and those dipped and subsequently annealed demonstrated less weight loss by the end of the study compared to untreated controls (8.25% vs. 7.05% average decrease in weight and 8.25% vs. 6.67%, respectively).

Banana Description	Day 1 (g)	Day 2 (g)	Day 3 (g)	Day 4 (g)	Day 5 (g)	Day 6 (g)	Day 7 (g)	Day 8 (g)	Δ in Weight (g)	Avg Δ in Weight (g)	% Δ in Weight	Avg % Δ in Weight
Control #1	38.27	36.64	35.19	34.25	33.18	31.95	30.43	29.12	-9.15	**-8.25**	-23.91%	**-23.39%**
Control #2	32.09	30.68	29.55	28.82	28.00	27.03	25.79	24.75	-7.34		-22.87%	
Dipped #1	36.27	35.06	34.00	33.27	32.44	31.53	30.40	29.43	-6.84	**-7.05**	-18.86%	**-19.26%**
Dipped #2	36.92	35.45	34.20	33.41	32.53	31.61	30.59	29.66	-7.26		-19.66%	
Dipped & Annealed #1	32.60	31.78	30.58	29.82	29.00	28.10	27.00	26.06	-6.54	**-6.67**	-20.06%	**-19.09%**
Dipped & Annealed #2	37.48	36.55	35.28	34.51	33.65	32.76	31.68	30.69	-6.79		-18.12%	

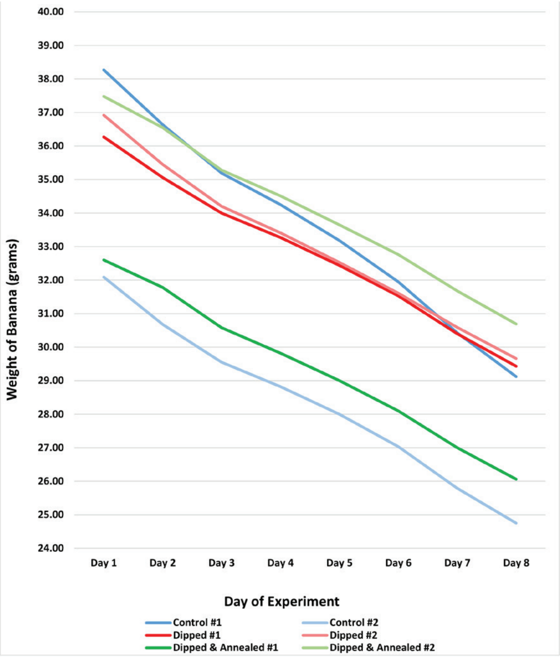


Figure 2: This graph displays the weight loss of each banana continuously from day 1 to day 8 of the study. Bananas in the experimental groups lost less weight (i.e. maintained freshness longer) than did the controls. This effect was most pronounced in the bananas that were dipped in silk fibroin solution and then water annealed.

Table 2: Reduction in weight of each banana after 8 days is shown. Untreated control bananas lost more weight by the end of the study than did either the dipped or the dipped and annealed fruit. The bananas that were dipped and subsequently water annealed lost the least amount of weight.

Banana Description	Loss in Weight (grams)
Control #1	9.15
Control #2	7.34
Dipped #1	6.84
Dipped #2	7.26
Dipped & Annealed #1	6.54
Dipped & Annealed #2	6.79

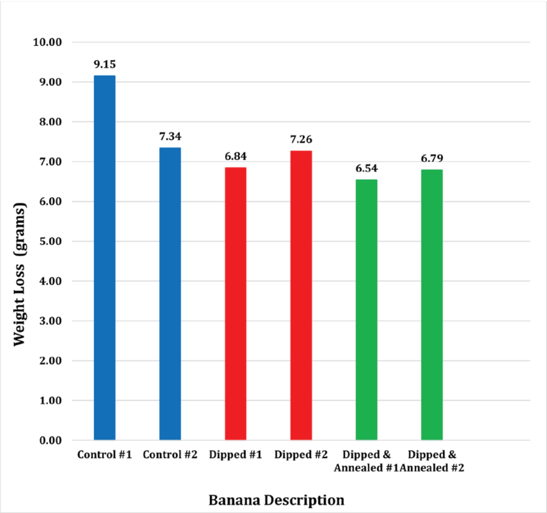


Figure 3: The bar graph depicts weight loss in each banana on day 8 of the study period. Weight loss, associated with fruit ripening and spoilage, was most pronounced in the control cohort. Bananas dipped in silk fibroin solution lost less weight than controls, and those dipped and subsequently annealed lost the least.

Table 3: Average percentage of weight loss in each cohort by day 8 is shown. Untreated control bananas lost an average 23.39% of their initial weight at the end of the study period, compared to 19.26% for those dipped in silk fibroin and 19.09% for the dipped and annealed cohort. Improved maintenance of weight corresponds to the preservation effect of the silk fibroin solution.

Bananas Description	Average Percent of Weight Loss (%)
Control #1 and #2	23.39%
Dipped #1 and #2	19.26%
Dipped & Annealed #1 and #2	19.09%

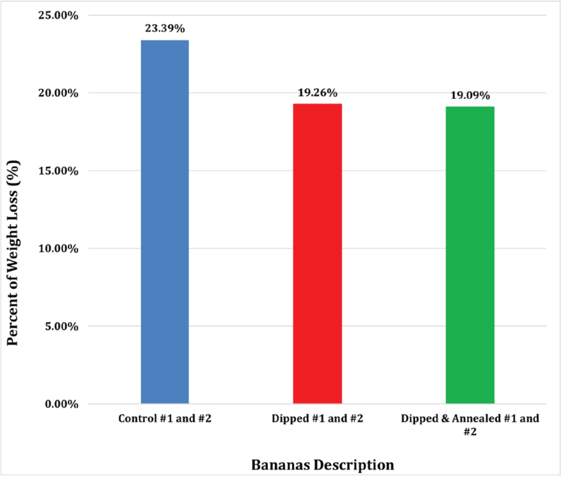


Figure 4: The bar graph depicts the average percentage of weight loss across all 3 cohorts (control, dipped, and dipped & annealed) after the 8 day study period. As expected, treated bananas maintained their initial weight better than untreated controls.

Turgidity of each banana was measured via the degree of depression, measured in millimeters, by the application of a 200 g weight. Given the volume loss of all bananas previously recorded, all experienced decreased turgidity after the study period, though in different amounts. The depth of depression by the weight for Control bananas 1 and 2 was 5 mm each; for Dipped bananas 1 and 2, the amount was 1.5 mm and 2 mm, respectively; and for Dipped & Annealed bananas 1 and

DOI: 10.36838/v8i2.69

2, the amount was 1 mm and 0.5 mm, respectively **(Table 4, Figure 5)**.

Table 4: The degree of vertical depression (as an indicator of turgor pressure) of each banana after 8 days via an applied 200 g weight is displayed. Turgidity was preserved in all 4 silk-fibroin treated bananas compared to controls, and this effect was most pronounced in the bananas dipped and then annealed. Untreated controls underwent ripening and softening sooner, resulting in an increased amount of depression with the applied weight.

Banana Description	Degree of Depression With 200 g Weight (mm)
Control #1	5.0
Control #2	5.0
Dipped #1	1.5
Dipped #2	2.0
Dipped & Annealed #1	1.0
Dipped & Annealed #2	0.5

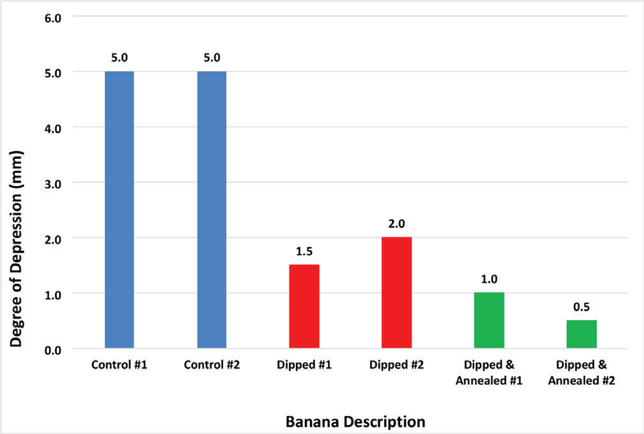


Figure 5: The graph demonstrates the degree of vertical depression of each banana on day 8 via an applied 200 g weight. The amount of depression was greatest in the control group, indicating loss of turgidity associated with ripening and spoilage.

To analyze the degree of microbial growth associated with fruit spoilage, mold/yeast colonies on agar plates were counted for each of the six bananas. Eight days after swabbing and plating, the agar for Control bananas 1 and 2 had 43 and 39 mold/yeast colonies, respectively, growing. Dipped bananas 1 and 2 showed 2 and 0 mold/yeast colonies, respectively. Dipped & Annealed bananas 1 and 2 showed 1 and 0 mold/yeast colonies, respectively. This data is presented in **Table 5, Figure 6**.

Table 5: Calculated mold and yeast colony counts grown on DRBC agar plates on days 1 to 8 of all bananas. Untreated controls showed the greatest degree of mold and yeast growth, associated with fruit spoilage. In comparison, silk fibroin dipped bananas grew significantly fewer organisms, and those dipped and annealed demonstrated the lowest number of colonies.

Banana Description	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8
Control #1	0	0	0	2	7	18	32	43
Control #2	0	0	0	2	4	7	20	39
Dipped #1	0	0	0	0	0	2	2	2
Dipped #2	0	0	0	0	0	0	0	0
Dipped & Annealed #1	0	0	0	0	0	1	1	1
Dipped & Annealed #2	0	0	0	0	0	0	0	0

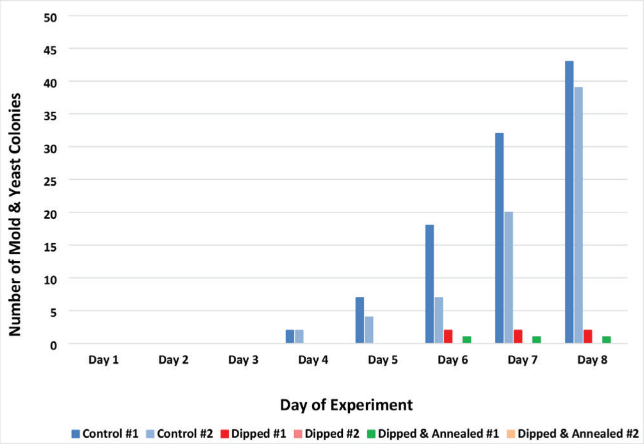


Figure 6: Mold and yeast colonies counted on DRBC agar for all bananas on days 1 to 8. Graphical representation of organism growth highlights the significant decrease in molds and yeast associated with fruit spoilage in the treated bananas compared to controls.

Discussion:

According to a 2021 report published by the U.S. Environmental Protection Agency,[1] approximately 35 to 36% of the U.S. food supply, or roughly 152 metric tons of food, is wasted along all the stages of the food supply chain. Notably, approximately 50% of this waste is experienced at the consumption portion of the supply chain. Fruits and vegetables, followed by dairy and eggs, are the most commonly wasted food items. Therefore, safe, effective, and readily available preservation methods for perishable foods are imperative.

Silk from *Bombyx mori* (silkworm) has been extensively studied in bioengineering due to its biocompatibility, robust mechanical performance, ease of processing, and ready supply.[5] Silk from silkworms is comprised of two primary proteins: sericin (25%) and fibroin (75%). Sericin is a glue-like, amorphous, and soluble protein that positions itself across the surface of two parallel fibroin fibers, binding them together and helping provide structural integrity of the fibers. Sericin can be removed from fibroin via a thermochemical process called degumming. Silk fibroin, in contrast, is a structural fibrous protein that adopts a semi-crystalline structure that imparts stiffness and strength. In its natural state, silk fibroin organizes into antiparallel beta-pleated sheets. A combination of beta-pleated sheets along with inter- and intra-molecular hydrogen bonding helps provide both flexibility and conformability.[2] These intrinsic properties of silk fibroin have allowed the protein to be used in a wide variety of biomedical applications, including drug delivery, biomaterial processing, wound healing, gene therapy, and bone regeneration.[2-5]

Notably, silk fibroin has also been studied in food processing as a way to preserve the post-harvest shelf life of perishable food. Silk fibroin has been investigated as a component in food packaging systems,[7,8] as well as an odorless and edible coating material on the foods themselves.[2,9] Prior studies have also demonstrated that the beta-pleated sheet content of the protein can be increased through a process called water annealing,[2,6] in which fruit coated with silk fibroin is exposed to water vapor in a vacuum-sealed environment and held at constant temperature for a period of time. Polar solvents favor the exposure of polar amino acid side chains to the solvent and bury nonpolar side chains within the core of the molecule, promoting folding of the protein. Additionally, beta-sheet structures are then stabilized by hydrogen bonding, increasing the formation and stability of this secondary structure. Increased beta-pleated sheet content correlated with improved food preservation.[2] This study sought to determine whether the application of a fibroin solution to fresh fruit would prolong its shelf life and reduce spoilage compared to an identical group of controls, and whether fibroin coating plus water annealing would be superior to coating alone.

Four parameters of fruit spoilage were utilized in our study: physical appearance, decrease in weight, loss of firmness/turgidity, and presence of microbial growth (specifically yeasts and molds). Digital photographs confirm decreased discoloration of the bananas treated with fibroin compared to controls, and the dipped and water annealed group looked fresher at the end of the period than did their dipped and non-annealed counterparts. After the 8-day banana ripening period, control bananas lost an average of 8.25 grams (or a 23.39% average decrease from their starting weight). Bananas that were dipped in fibroin lost less weight than controls: an average of 7.05 grams or 19.26% average decrease. The bananas that were both dipped and then water annealed for 12 hours lost the least amount of weight: 6.67 grams (19.07%). Loss of fruit firmness was documented by the amount of compression caused by application of a 200-gram weight, and a similar trend was seen: control bananas were least firm, with a compression height of 5 mm; dipped bananas compressed between 1.5 and 2 mm, and dipped & annealed bananas showed only 0.5 to 1 mm of compression. The most common organisms associated with fruit spoilage are molds and yeasts. Therefore, dichloran rose bengal chloramphenicol (DRBC) was utilized in the microbial analysis in this study because of the medium's superiority in enumerating foodborne yeasts and molds, specifically.[10] Although the bananas that were dipped and annealed showed modest improvements across all measured parameters compared to their dipped-only counterpoints, the small sample size in the study precludes the ability to say whether these differences reached statistical significance.

These results compare favorably to a prior study[2] that compared ripening of fresh strawberries and bananas that were fibroin dipped, dipped and water annealed, and non-dipped. The investigators noted statistically significant improvement in water loss, weight loss, and oxygen diffusion in dipped versus control strawberries; the effect was more pronounced in the annealed cohort. Similarly, treated bananas better maintained turgidity and firmness in a statistically significant fashion than did untreated controls.

Limitations of this study included a small sample size. A larger number of bananas used in the study would have added statistical power to the investigation and allowed for an analysis of whether measured differences between groups were statistically significant. Additionally, quantitative analysis using instrumentation to assess textural and color differences between bananas and quantifying ethylene gas influx as a marker for browning seen in ripe bananas would reduce some of the more subjective assessments that were made in the study. These represent areas of research for future studies.

Conclusion

A 1% weight/volume aqueous silk fibroin solution effectively extended the post-harvest shelf life of bananas in all of the parameters measured. This effect was even more pronounced in the fruit that was coated and water annealed. It is thought that the crystalline beta-pleated sheet coating of the hydrophobic fibroin prolongs the freshness of fruit by slowing fruit respiration, decreasing water loss and dehydration, and subsequently extending fruit firmness. Water annealing increases the degree of beta-pleated sheet content of the fibroin coating, thereby enhancing its preservative effect. The implication of this study, and that of hopeful future follow-up investigations, would be the addition of silk fibroin as a safe, effective, and readily available preservative to the armamentarium against the staggering effects of food waste.

Acknowledgments

I would like to express my sincerest gratitude to Mr. Wenzen Chuang, my research mentor. His expertise, encouragement, and guidance were instrumental in the development and completion of this project.

References

1. U.S. EPA (U.S. Environmental Protection Agency). (2021). Part 1, From Farm to Kitchen: The Environmental Impacts of U.S. Food Waste. EPA 600-R21 171. https://www.epa.gov/system/files/documents/2021-11/from-farm-to-kitchen-the-environmental-impacts-of-u.s.-food-waste_508-tagged.pdf
2. Marelli, B., Brenckle, MA, Kaplan, DL, Omenetto, FG. Silk Fibroin as Edible Coating for Perishable Food Preservation. Sci Rep. 2016 May 6;6:25263. Doi: 10.1038/srep25263. PMID: 27151492; PMCID: PMC4858704.
3. Suzuki Y. Structures of Silk Fibroin Before and After Spinning and Biomedical Applications. Polymer Journal. 2016;48:1039-1044. doi: 10.1038/pj.2016.77.
4. Nguyen TP, Nguyen QV, Nguyen VH, Le TH, Huynh VQN, Vo DN, Trinh QT, Kim SY, Le QV. Silk Fibroin-Based Biomaterials for Biomedical Applications: A Review. Polymers (Basel). 2019 Nov 24;11(12):1933. doi: 10.3390/polym11121933. PMID: 31771251; PMCID: PMC6960760.
5. Qi Y, Wang H, Wei K, Yang Y, Zheng RY, Kim IS, Zhang KQ. A Review of Structure Construction of Silk Fibroin Biomaterials from Single Structures to Multi-Level Structures. Int J Mol Sci. 2017 Mar 3;18(3):237. doi: 10.3390/ijms18030237. PMID: 28273799; PMCID: PMC5372488.
6. Hu X, Shmelev K, Sun L, Gil ES, Park SH, Cebe P, Kaplan DL. Regulation of silk material structure by temperature-controlled water vapor annealing. Biomacromolecules. 2011 May 9;12(5):1686-96. doi: 10.1021/bm200062a. Epub 2011 Mar 22. PMID: 21425769; PMCID: PMC3090511.
7. Valentini L, Bittolo Bon S, Pugno NM. Combining Living Microorganisms with Regenerated Silk Provides Nanofibril-Based Thin Films with Heat-Responsive Wrinkled States for Smart Food Packaging. Nanomaterials (Basel). 2018 Jul 11;8(7):518. Doi: 10.3390/nano8070518. PMID: 29997336; PMCID: PMC6071141.

8. Giannelli M, Lacivita V, Posati T, Aluigi A, Conte A, Zamboni R, Del Nobile MA. Silk Fibroin and Pomegranate By-Products to Develop Sustainable Active Pad for Food Packaging Applications. Foods. 2021 Nov 25;10(12):2921. doi: 10.3390/foods10122921. PMID: 34945471; PMCID: PMC8700627.
9. Muthumanickam, Muthulakshmi & Palanivel, Rameshthangam & Ambiga, C & Sindhamani, S & Ramya, R & Gomathirajashyamala, L. (2024). Application of Silk Fibroin Nanoparticles Based Edible Coating Material for Postharvest, Shelf-Life Extension and Preservation of Food Products. Shanlax International Journal of Arts, Science and Humanities. 12. 180-193. 10.34293/sijash.v12i1.8029.
10. Beuchat LR, Mann DA. Comparison of New and Traditional Culture-Dependent Media for Enumerating Foodborne Yeasts and Molds. J Food Prot. 2016 Jan;79(1):95-111. doi: 10.4315/0362-028X.JFP-15-357. PMID: 26735035

Author

Matthew Martinelli is a junior at Highland Park High School in Texas. His dedication to community service has inspired several research projects and publications investigating issues from food preservation to the health of agricultural workers. He is currently assisting a professor in a research lab at Southern Methodist University.

IJHSR ijhighschoolresearch.org

■ REVIEW ARTICLE

Innovations in Skin Regeneration: The Intersection of 3D Bioprinting and Single-Cell RNA Sequencing

Jihoo Hyun

Canyon Crest Academy, 5951 Village Center Loop Rd, San Diego, CA 92130, USA; 0903stella@gmail.com

ABSTRACT: The combination of single-cell RNA sequencing (scRNA-seq) and 3D bioprinting technologies is transforming the field of skin tissue engineering by providing unmatched levels of accuracy and control at the cellular level. This article is on the revolutionary developments in skin regeneration with these technologies. Although 3D bioprinting allows one to print intricate tissue architecture similar to native skin, it also creates challenges of scalability and proper vascularization. scRNA-seq overcomes these limitations by offering high-resolution information regarding cellular heterogeneity and gene expression patterns in bioprinted tissues. Merging these technologies, scientists can engineer bioprinting approaches and create more proficient biomaterials to finally boost the regenerative potential and function of fabricated skin tissues. This review offers a systematic overview of the latest research, emphasizing the synergy between 3D bioprinting and scRNA-seq, with their respective contribution to the improvement of skin regeneration and the development of more accurate and effective therapeutic strategies.

KEYWORDS: Biomedical Engineering, Biomaterials and Regenerative Medicine, Skin Regeneration, 3D Bioprinting, Single-Cell RNA Sequencing.

■ Introduction

Skin regeneration is an important field of regenerative medicine, where technology plays a key role in enhanced clinical outcomes in the event of burns, chronic wounds, or trauma.[1] As the largest organ and the initial barrier against insults of the external environment, skin is a complex, multilayered tissue with heterogeneous cell populations and extracellular components having specific biological functions. Rebuilding the architecture and functionality of damaged skin is challenging, and traditional grafts or scaffolds are insufficient to meet these demands.[2]

The breakthroughs in biomedical engineering have presented us with two technologically powerful weapons for overcoming these challenges: single-cell RNA sequencing (scRNA-seq) and 3D bioprinting. Both emergent technologies are rapidly transforming skin tissue engineering with mutually complementary advantages. scRNA-seq enables one to perform high-resolution transcriptomic analysis on the cellular level, which enables researchers to reveal cellular heterogeneity, gene expression variation, and differentiation programs with unprecedented accuracy.[3] This information is crucial to fine-tune cell selection and molecular target identification to guide tissue regeneration.

Simultaneously, 3D bioprinting provides the spatially controlled bioprinting of living cells and biomaterials to produce structured skin constructs with both structural and functional homology to native tissue. By virtue of the capability to model the epidermal, dermal, and hypodermal stratified structure, 3D bioprinting makes it possible to create tissue models addressed to specific patient requirements.[4] Combining scRNA-seq and 3D bioprinting offers unparalleled potential to boost the biological fidelity, integration, and therapeutic potency of engineered skin tissues.[5]

As highlighted by Murphy and Atala,[4] 3D bioprinting has already demonstrated success in constructing anatomically precise and viable tissue analogues, setting the stage for broader clinical applications. Moreover, Farage *et al.* [3] emphasize that both intrinsic and extrinsic factors, such as aging, environmental exposure, and systemic conditions, profoundly influence skin regeneration, further validating the need for advanced approaches that integrate both structural and molecular considerations. The convergence of these technologies marks a new era in skin regeneration, bringing us closer to developing functional, patient-specific skin grafts for real-world therapeutic use.

■ Literature Review Approach

This review synthesizes existing literature on 3D bioprinting and single-cell RNA sequencing in the context of skin tissue engineering. It includes peer-reviewed journal articles from fields such as tissue engineering, regenerative biology, and bioinformatics. Key studies were selected based on relevance, methodological rigor, and impact on the field. Emphasis was placed on studies integrating both bioprinting techniques and transcriptomic analysis. No direct experimental work was conducted; instead, a comprehensive literature analysis approach was used.

■ Advances in Skin Regeneration Technologies

Structure and Function of Human Skin:

The barrier function of the skin is intimately associated with its stratified nature, in which each stratum has a specific role in its protective capabilities. The epidermis, made up primarily of keratinocytes, is a hard, impermeable barrier to water that

DOI: 10.36838/v8i2.75

guards against dehydration and invasion by microorganisms. Injury to this stratum jeopardizes the protective capability of the skin and makes it vulnerable to infection. The dermis, located beneath the epidermis, contains collagen and elastin and is responsible for strength and elasticity, and contains vital structures such as blood vessels, nerve endings, hair follicles, and sweat glands. Injury to the dermis, due to second-degree burns, can destabilize these structures and affect functions such as thermoregulation, regulation of moisture, and sensation. The hypodermis, which is the innermost layer, is a store of energy, an insulator, and a shock absorber. Injury to this layer compromises the body's ability to regulate temperature and cushion against physical trauma.[1] Additionally, deeper wounds can damage nerves and blood vessels, causing pain, numbness, and impaired circulation.[2] Figure 1 illustrates the layered structure of the skin and its key anatomical features involved in protection and function.

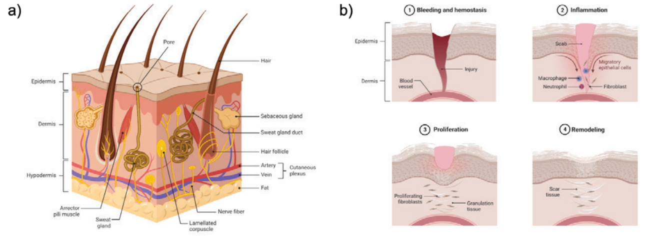


Figure 1: The body's largest organ, the skin, is made up of three main layers: the epidermis, dermis, and hypodermis. The outer layer, the epidermis, creates our color and a water-repellent barrier. The layer underneath the epidermis, the dermis, is made up of hard connective tissue, hair follicles, and sweat glands. The bottom layer, the hypodermis, is made up of fat and connective tissue. These layers combined shield the body from environmental damage, pathogens, and physical trauma. [6]

Healing Process and Regeneration Challenges:

The skin's healing process involves four overlapping stages: hemostasis, inflammation, proliferation, and remodeling. Hemostasis begins immediately after trauma, with the clotting of blood to prevent unnecessary loss of blood. Next, there is inflammation, where the immune cells clean out the debris and combat any infection. During the proliferation phase, new tissue formation takes place, including the regeneration of the epidermis and collagen deposition in the dermis. Remodeling finally tightens the new tissue and adapts it for function.[7]

Serious injury presents enormous difficulties for this process. Severe damage has the potential to create faulty regeneration, chronic wounds, or scarring, which is not as fully functional as normal skin. And illnesses like impaired blood supply, infection, and recurrent trauma can also delay healing. 1 Diabetic ulcers, for instance, may not move beyond the healthy stages of healing, stuck in an extended inflammatory phase that does not allow tissue formation.[8] And besides, large wounds need massive amounts of tissue formation, which the body cannot provide on its own in adequate amounts, resulting in fibrous scar tissue that doesn't have the same elasticity or tensile strength as healthy skin.[9]

Clinical Importance of Skin Regeneration:

Skin is the body's largest organ, serving as a crucial barrier against environmental damage, pathogens, and physical injuries. Effective skin regeneration is vital for treating burns, chronic wounds, and other skin-related conditions. The ability to restore the skin's structure and function can significantly improve patients' quality of life and reduce healthcare costs associated with long-term wound care. Seven million people worldwide suffer from severe skin injuries each year. For example, the World Health Organization (WHO) estimates that burns alone cause approximately 180,000 deaths annually, with the majority occurring in low- and middle-income countries. Additionally, chronic wounds such as diabetic ulcers affect millions more, leading to prolonged hospital stays, frequent medical interventions, and a substantial financial burden on healthcare systems.[10]

■ 3D Bioprinting in Skin Tissue Engineering

Overview and Techniques:

3D bioprinting represents an emerging method of additive manufacturing that enables the precise production of tissue constructs through sequential deposition of bioinks, often consisting of living cells, growth factors, and biomaterials. A computer-aided design model instructs the spatial organization of bioinks in three-dimensional space to allow for the fabrication of intricate tissue structures mimicking hierarchically organized skin units in native skin. In skin tissue engineering, bioprinting can produce stratified layers that resemble the epidermis, dermis, and hypodermis, incorporating various cell types such as keratinocytes, fibroblasts, and melanocytes in their respective layers. These tissues may also be seeded with extracellular matrix (ECM) components such as collagen, hyaluronic acid, and fibrin to enhance the structural integrity and cell-cell interactions. 3D bioprinting by incorporating fine control of cell and biomaterial deposition provides the capability of tailoring tissue properties such as porosity, stiffness, and biochemical content, which are essential in replicating the functional and mechanical characteristics of native skin. Figure 2 outlines the key steps of the bioprinting workflow and illustrates various bioprinting techniques used to fabricate engineered tissues.

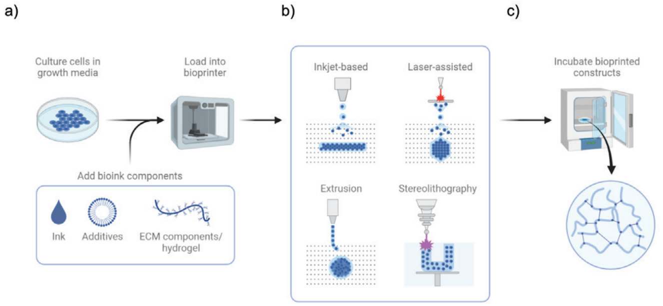


Figure 2: The process involves the use of various bioprinting techniques, such as inkjet bioprinting, extrusion bioprinting, and laser-assisted bioprinting, each with distinct advantages and limitations in terms of resolution, cell viability, and material compatibility.[4,11] Inkjet bioprinting uses thermal or acoustic forces to deposit droplets of bioink, making it suitable for high-resolution printing of delicate structures. Extrusion bioprinting, which forces bioink through a nozzle, allows for continuous deposition and is ideal for constructing larger tissue volumes. Laser-assisted bioprinting employs laser energy to propel cells and biomaterials onto a substrate, providing high precision and control over the placement of individual cells.

Recent trends in 3D bioprinting include multi-material printing, which enables the incorporation of different cell types and materials within a single construct to better mimic the natural tissue environment. This approach is particularly useful for creating skin tissues, which consist of multiple layers with distinct cellular compositions and functions. Advances in bioprinting also focus on enhancing vascularization within the printed tissues to ensure proper nutrient and oxygen supply, a critical factor for the viability and function of complex tissue constructs.[12] The use of bioactive materials that promote cell growth and differentiation is another important trend, as these materials can significantly improve the regenerative potential of bioprinted tissues.[13]

Innovations and Clinical Applications:

Precedents for the application of 3D bioprinting in tissue engineering include studies where bioprinting has been used to create complex structures such as vascularized tissues, cartilage, and bone. For instance, Homan *et al*. demonstrated the bioprinting of 3D renal structures that mimicked the function and structure of native kidney tissues.[16] This study highlights the potential of bioprinting to create functionally relevant tissue constructs. Similarly, Daly *et al*. developed bioprinted cartilage constructs that closely resembled the mechanical properties and cellular organization of natural cartilage, showcasing the versatility of bioprinting in generating various tissue types.[17] Moreover, Zhang *et al*. successfully created bioprinted bone tissues with integrated vascular networks, emphasizing the importance of vascularization for the survival and integration of bioprinted tissues.[18]

Studies have demonstrated the potential of 3D bioprinted tissues to replicate the structural and functional properties of native skin, making them suitable for clinical applications. For example, Kador *et al*. combined 3D printing with radial electrospun scaffolds to control retinal ganglion cell positioning and neurite growth, demonstrating the precision of bioprinting in creating structured tissue environments.[19] Intini *et al*. explored the use of 3D-printed chitosan-based scaffolds for skin regeneration, finding that these scaffolds supported skin cell growth and wound healing effectively.[20] Kandarova and Hayden utilized standardized reconstructed skin models to study cellular responses to different bioprinting strategies, providing detailed insights into the optimization of bioprinting techniques for skin tissue engineering.[21]

Bioactive Materials and Multi-Material Printing:

The development of bioactive materials and the incorporation of multiple cell types within bioprinted constructs represent significant advancements in the field. Zhu *et al*. highlighted the use of bioinks containing growth factors and ECM components to enhance cell proliferation and differentiation, improving the overall functionality of bioprinted tissues.[22] The use of multi-material printing techniques enables the creation of more complex tissue architectures, better replicating the natural environment of skin tissues and enhancing their regenerative potential.

The rapid advancements in 3D bioprinting for skin tissue engineering have been tempered by persistent challenges that limit its full potential, such as achieving effective vascularization, scalability, and the precision needed to replicate the complex architecture of natural skin.[12,13] Additionally, the standardization of bioinks remains an obstacle.[14] These limitations underscore the need for complementary technologies like single-cell RNA sequencing (scRNA-seq), which provides a detailed understanding of cellular heterogeneity and gene expression dynamics at the single-cell level. By integrating scRNA-seq with 3D bioprinting, researchers can refine tissue constructs, ensuring that the cellular composition and function closely mirror those of native tissues.[3,15] This integration not only enhances the reproducibility and functionality of bioprinted tissues but also opens new avenues for the development of more effective and clinically applicable skin regeneration therapies.

■ Integration of Single-Cell RNA Sequencing (scRNA-seq)

Role in Analyzing Bioprinted Tissues:

Single-cell RNA sequencing (scRNA-seq) provides high-resolution insights into the transcriptomic landscape of individual cells, making it a powerful tool for understanding cellular heterogeneity, identifying distinct cell populations, and tracking gene expression changes during tissue regeneration. Figure 3 illustrates the key steps involved in scRNA-seq, from tissue dissection to sequencing and cell type identification, highlighting its utility in regenerative studies.

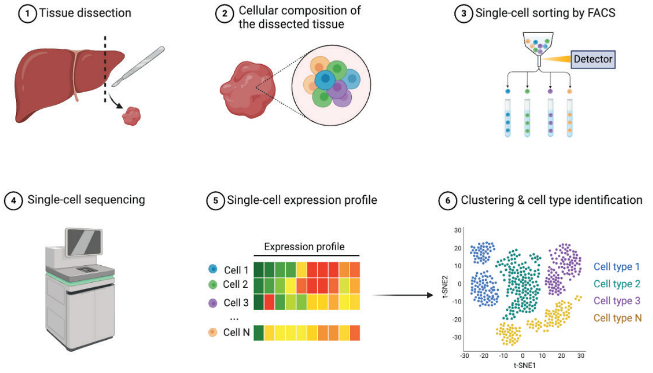


Figure 3: This technology involves isolating individual cells, reverse transcribing their RNA into complementary DNA (cDNA), and sequencing the cDNA to obtain detailed gene expression profiles. The resulting data can be used to construct a comprehensive map of cellular states and interactions within a tissue, revealing the molecular mechanisms underlying tissue regeneration and repair. [3,15]

Applications in Regenerative Studies:

As the field of skin tissue engineering advances, it becomes clear that integrating innovative technologies is essential for overcoming current limitations and enhancing regeneration outcomes. Despite the significant progress made with 3D bioprinting, challenges such as scalability, precision, and effective vascularization persist. This is where single-cell RNA sequencing (scRNA-seq) emerges as a transformative tool. By providing high-resolution insights into the transcriptomic landscape of individual cells, scRNA-seq allows researchers to analyze the cellular composition and gene expression profiles of bioprinted tissues. This integration supports the validation and optimization of bioprinting strategies, leading to more functional and reliable skin constructs. The following sections highlight various studies that demonstrate the synergy between scRNA-seq and 3D bioprinting in the context of skin regeneration.

Use Cases in Skin Regeneration:

Precedents for using scRNA-seq in similar contexts include studies on the regenerative processes in other tissues and organs. For instance, scRNA-seq has been used to profile the cellular landscape of regenerating heart tissue, providing insights into the roles of various cell types during cardiac repair.[23] Similarly, scRNA-seq has been employed to study the cellular dynamics in regenerating liver tissue, identifying key regulatory genes and pathways involved in liver regeneration.[24] In the field of neural tissue engineering, scRNA-seq has helped uncover the heterogeneity of neural stem cells and their differentiation pathways, facilitating the development of more effective strategies for neural regeneration.[25]

In the context of skin tissue engineering, scRNA-seq is particularly valuable for analyzing the cellular composition and gene expression profiles of bioprinted tissues. By comparing these profiles to those of native skin, researchers can assess the degree of similarity and identify areas for improvement in the bioprinting process. This approach has been used to validate the efficacy of various bioprinting strategies and biomaterials, providing crucial information for optimizing tissue engineering techniques.[3,15] For example, scRNA-seq can reveal differences in the expression of key genes involved in skin development, inflammation, and wound healing, allowing researchers to fine-tune the bioprinting parameters and improve the functional integration of bioprinted tissues. Solé-Boldo *et al*. demonstrated the application of scRNA-seq in understanding stem cell heterogeneity within engineered tissues, aiding in the optimization of cellular compositions and functional outcomes in bioprinted constructs.[26] Additionally, Guo *et al*. utilized scRNA-seq to evaluate cellular dynamics in skin regeneration, revealing key pathways that contribute to tissue integration and repair.[27] Similarly, Tabib *et al*. applied scRNA-seq to assess immune cell infiltration in bioprinted skin models, leading to improved protocols for creating more immunocompatible tissue constructs.[15]

Another notable application is in the study of skin aging, where scRNA-seq has been used to identify age-related changes in cellular composition and gene expression in human skin. This research has provided valuable insights into the molecular mechanisms of skin aging and potential targets for rejuvenation therapies.[26] Additionally, scRNA-seq has been utilized to investigate the immune response in wound healing, revealing the roles of various immune cell populations in tissue repair and the resolution of inflammation.[27]

Integration of scRNA-seq and 3D Bioprinting in Skin Tissue Engineering:

Building on the advancements discussed in enhancing regenerative potential, studies exemplifying the use of single-cell RNA sequencing (scRNA-seq) and 3D bioprinting in skin tissue engineering are critical for advancing the field. These studies underscore the importance of effective skin regeneration in medical treatments and wound healing. Intini *et al*. revealed that chitosan-based scaffolds support skin cell growth, with scRNA-seq confirming their alignment with chitosan regenerative properties.[20] Farage *et al*. demonstrated that silk fibroin hydrogels, analyzed via scRNA-seq, promote scarless skin regeneration by recruiting specific cell populations.[3] Additionally, Tabib *et al*. used scRNA-seq to identify age-related declines in dermal sheath cells, suggesting potential rejuvenation strategies.[15] Kandarova and Hayden applied scRNA-seq to optimize bioprinted skin models, enhancing the precision of bioprinting techniques.[21] Kolesky *et al*. validated the integration of vascular networks in bioprinted tissues through scRNA-seq, emphasizing vascularization's role in tissue viability. [12] Finally, Kim *et al*. highlighted the optimization of bioinks that mimic the extracellular matrix, with scRNA-seq confirming their role in promoting tissue regeneration.[5] Collectively, these studies demonstrate the pivotal role of integrating scRNA-seq with 3D bioprinting to advance skin tissue engineering, improving outcomes in addressing severe burns, chronic wounds, and other skin-related medical conditions.

■ Current Challenges and Future Perspectives

Despite the promising results, several challenges remain. These include the scalability of bioprinted tissues, the standardization of bioinks, and the need for comprehensive bioinformatics tools to analyze scRNA-seq data.[28,29] Advanced transcriptomic techniques, such as spatial transcriptomics and multi-omics approaches, are being integrated to provide more precise and context-aware insights into cellular behavior within bioprinted tissues. For example, spatial transcriptomics allows for the mapping of gene expression within the spatial architecture of tissues, providing valuable information on how different cell types organize and function within engineered constructs.[30] Additionally, single-cell multi-omics approaches combine scRNA-seq with other modalities like epigenomics and proteomics to capture a more comprehensive profile of cellular states and interactions.[31] The current limitations in the resolution and precision of bioprinting techniques need to be addressed to create more complex and functional tissue constructs. Furthermore, the integration of advanced bioinformatics tools is essential for managing and interpreting the vast amounts of data generated by scRNA-seq and related methods, which can provide deeper insights into the cellular processes underlying tissue regeneration. Future research should focus on developing advanced bioinks that mimic the ECM, optimizing bioprinting techniques to enhance cell viability, and employing scRNA-seq, spatial transcriptomics, and multi-omics approaches to continuously validate and refine tissue engineering approaches.[5,32]

Conclusion and Outlook

The marriage of 3D bioprinting and single-cell RNA sequencing (scRNA-seq) represents an advancement in skin tissue engineering. These novel technologies share complementary strengths that overcome the shortcomings of conventional techniques and maximize the regenerative capabilities of engineered skin tissues. 3D bioprinting is renowned for having the strength of spatial structural control, the potential to develop complex tissue architecture resembling the native skin structure in its details. But it has remained confronted by limitations such as scalability, vascularization, and standardization of the material.

scRNA-seq overcomes these limitations by achieving high-resolution access to molecular and cellular dynamics during tissue regeneration. By comparing bioprinted tissue cell composition and gene expression profiles, scRNA-seq enables us to tailor bioprinting approaches and create more effective biomaterial designs. This coexistence broadens our knowledge of skin regeneration and makes it possible to generate more functional and realistic skin constructions.

With research ongoing, the future progression and intersection of scRNA-seq and 3D bioprinting will be crucial to regenerative medicine advancement. The two technologies possess enormous potential for expanding patient advantages as well as developing new, more effective treatments for skin repair and regeneration, guiding the future of tissue engineering.

Acknowledgments

I would also like to express my gratitude to my Stanford University mentor, whose support, guidance, and critical remarks were vital in directing this research. His regenerative medicine and tissue engineering knowledge was invaluable in helping me articulate my ideas and significantly added value to the quality of this review.

References

1. Singer, A. J., & Clark, R. A. (1999). Cutaneous wound healing. New England Journal of Medicine, 341(10), 738–746.
2. Tiwari, V. K. (2012). Burn wound: How does it differ from other wounds? Indian Journal of Plastic Surgery, 45(2), 364–373.
3. Farage, M. A., Miller, K. W., Elsner, P., & Maibach, H. I. (2008). Intrinsic and extrinsic factors in skin ageing: A review. International Journal of Cosmetic Science, 30(2), 87–95.
4. Murphy, S. V., & Atala, A. (2014). 3D bioprinting of tissues and organs. Nature Biotechnology, 32(8), 773–785.
5. Kim, G., Ahn, S., Yoon, H., Kim, Y., & Chun, W. (2009). A cryogenic direct-plotting system for fabrication of 3D collagen scaffolds for tissue engineering. Journal of Materials Chemistry, 19(46), 8817–8823.
6. Proksch, E., Brandner, J. M., & Jensen, J. M. (2008). The skin: An indispensable barrier. Experimental Dermatology, 17(12), 1063–1072.
7. Gurtner, G. C., Werner, S., Barrandon, Y., & Longaker, M. T. (2008). Wound repair and regeneration. Nature, 453(7193), 314–321.
8. Falanga, V. (2005). Wound healing and its impairment in the diabetic foot. The Lancet, 366(9498), 1736–1743.
9. Broughton, G., Janis, J. E., & Attinger, C. E. (2006). The basic science of wound healing. Plastic and Reconstructive Surgery, 117(7S), 12S–34S.
10. Sen, C. K., Gordillo, G. M., Roy, S., *et al.* (2009). Human skin wounds: A major and snowballing threat to public health and the economy. Wound Repair and Regeneration, 17(6), 763–771.
11. Groll, J., Boland, T., Blunk, T., *et al.* (2016). Biofabrication: Reappraising the definition of an evolving field. Biofabrication, 8(1), 013001.
12. Kolesky, D. B., Homan, K. A., Skylar-Scott, M. A., & Lewis, J. A. (2016). Three-dimensional bioprinting of thick vascularized tissues. Proceedings of the National Academy of Sciences, 113(12), 3179–3184.
13. Zhang, Y. S., Arneri, A., Bersini, S., *et al.* (2017). Bioprinting 3D microfibrous scaffolds for engineering endothelialized myocardium and heart-on-a-chip platforms. Biomaterials, 110, 45–59.
14. Zhu, W., Qu, X., Zhu, J., *et al.* (2016). Direct 3D bioprinting of prevascularized tissue constructs with complex microarchitecture. Biomaterials, 124, 106–115.
15. Tabib, T., Morse, C., Wang, T., *et al.* (2018). Single-cell RNA-seq identifies pathogenic fibroblasts in autoimmune disease. Nature, 563, 229–233.
16. Homan, K. A., Kolesky, D. B., Skylar-Scott, M. A., *et al.* (2016). Bioprinting of 3D convoluted renal proximal tubules on perfusable chips. Scientific Reports, 6, 34845.
17. Daly, A. C., Cunniffe, G. M., Sathy, B. N., *et al.* (2016). 3D bioprinting of developmentally inspired templates for whole bone organ engineering. Advanced Healthcare Materials, 5(18), 2353–2361.
18. Zhang, Y. S., Arneri, A., Bersini, S., *et al.* (2017). Bioprinting 3D microfibrous scaffolds for engineering endothelialized myocardium and heart-on-a-chip platforms. Biomaterials, 110, 45–59.
19. Kador, K. E., Grogan, S. P., Dorthé, E. W., *et al.* (2016). Control of retinal ganglion cell positioning and neurite growth: Combining 3D printing with radial electrospun scaffolds. Tissue Engineering Part A, 22(3–4), 286–294.
20. Intini, C., Elviri, L., Cabral, J., *et al.* (2018). 3D-printed chitosan-based scaffolds: An in vitro study of human skin cell growth and an in vivo wound healing evaluation in experimental diabetes in rats. Carbohydrate Polymers, 199, 593–602.
21. Kandarova, H., & Hayden, P. J. (2021). Standardised reconstructed skin models in toxicology and pharmacology: State of the art and future development. Handbook of Experimental Pharmacology, 265, 57–71.
22. Zhu, W., Qu, X., Zhu, J., *et al.* (2016). Direct 3D bioprinting of prevascularized tissue constructs with complex microarchitecture. Biomaterials, 124, 106–115.
23. Sereti, K. I., Zhang, Y., Arneri, A., *et al.* (2018). Single-cell RNA sequencing of regenerating heart tissue reveals heterogeneity in cell state and lineage relationships. Nature Communications, 9, 4436.
24. Halpern, K. B., Shenhav, R., Matcovitch-Natan, O., *et al.* (2017). Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. Nature, 542, 352–356.
25. Zeisel, A., Hochgerner, H., Lönnerberg, P., *et al.* (2018). Molecular architecture of the mouse nervous system. Cell, 174(4), 999–1014.
26. Solé-Boldo, L., Raddatz, G., Schütz, S., *et al.* (2020). Single-cell transcriptomes of the human skin reveal age-related loss of fibroblast priming. Nature Aging, 1, 123–135.
27. Guo, X., Zhang, Y., Zheng, L., *et al.* (2018). Single-cell RNA-seq reveals the temporal diversity and dynamics of immune cells during acute inflammation. Cell Reports, 25(4), 1125–1136.

DOI: 10.36838/v8i2.75

28. Varani, J., Dame, M. K., Rittie, L., *et al*. (2006). Decreased collagen production in chronologically aged skin. The American Journal of Pathology, 168(6), 1861–1868.
29. Zou, Z., Long, X., Zhao, Q., *et al*. (2021). Ageroprotective effect of the HES1 transcription factor. Nature Communications, 12(1), 4412.
30. Philippeos, C., Telerman, S. B., Oulès, B., *et al*. (2018). Spatial and single-cell transcriptional profiling identifies functionally distinct human dermal fibroblast subpopulations. The Journal of Clinical Investigation, 128(4), 1780–1795.
31. Keriquel, V., Oliveira, H., Rémy, M., *et al*. (2017). In situ printing of mesenchymal stromal cells, by laser-assisted bioprinting, for in vivo bone regeneration applications. Scientific Reports, 7(1), 1778.
32. Komez, A., Yildiz, A., & Ozcelik, B. (2016). Construction of a patterned hydrogel-fibrous mat bilayer structure to enhance skin regeneration. Journal of Materials Chemistry B, 4(15), 2550–2558.

■ Authors

Jihoo Hyun is a junior at Canyon Crest Academy with interests in genetics, tissue engineering, and bioinformatics. She hopes to pursue biomedical research and has participated in multiple biology and computer science programs.

IJHSR

ijhighschoolresearch.org

■ RESEARCH ARTICLE

Impact of Physiological Stress on Decision Accuracy among Elite Chess Players: A Biometric Analysis

Yash Jayesh Laddha[1], Shubh Jayesh Laddha[2]

1. Greenwood High International School, No. 8-14 Chikkawadayarapura, Bangalore, Karnataka, 560087, India; yashladdha75@gmail.com
2. Delhi Public School East, Survey No. 43, Dommasandra Post, Bangalore, Karnataka, 562125, India

ABSTRACT: The impact of physiological stress on cognitive performance in high-pressure environments remains a relatively underexplored area of research, particularly in settings such as chess, where decision-making is rapid and cognitively demanding. Although heart rate (HR) and cognitive stress responses have been studied in sports and clinical settings, few studies have examined their impact in elite-level mental competitions. This study aimed to investigate whether elevated HR and time pressure (<30 seconds remaining) negatively affect decision-making accuracy in professional chess players and whether this relationship was influenced by age. We analyzed biometric and performance data from 50 publicly available blitz chess games played by 50 grandmasters. HR data was collected at regular intervals, and move accuracy was calculated using Chess.com's evaluation system. We found that players with higher average HRs (>130 bpm) played with significantly lower accuracy, especially under time pressure. A negative correlation and regression model further confirmed that HR was a significant predictor of accuracy. Additionally, older players had higher HRs under stress. These findings provide novel evidence of how physiological stress can negatively impact mental performance and have important implications for chess players since they highlight the potential of biofeedback-based training to improve decision-making in high-pressure scenarios.

KEYWORDS: Behavioral and Social Sciences, Physiological Psychology, Cognitive Stress, Heart Rate, Chess Performance Analysis.

Introduction

High-pressure environments consistently elicit physiological stress responses that can impact human cognition.[1] In such contexts, decision-making under pressure is closely connected with biological arousal systems.[2,3] Chess, while traditionally focused on cognitive skills and tactical ability, offers a unique platform for studying this relationship. Its demands on memory, attention, and problem-solving make it ideal for analyzing the effects of stress on decision quality.[4]

Among the most widely studied physiological markers of stress are heart rate (HR) and heart rate variability (HRV), which reflect sympathetic and parasympathetic nervous system activation, respectively.[5] Elevated HR is an indicator of sympathetic arousal, often accompanied by mental and acute stress.[6] HRV, particularly the vagally mediated high-frequency component, reflects autonomic flexibility and is inversely related to stress reactivity.[7,8] Research across cognitive, clinical, and occupational settings has shown that lower HRV correlates with impaired decision-making, reduced attentional control, and poorer performance during high-demand tasks. Conversely, individuals with higher resting HRV exhibit better self-regulation, faster recovery from stress, and improved decision-making accuracy under tension.[5,9]

This relationship becomes significant in high-performance environments.[6] For example, first responders with low HRV have been shown to underperform in simulated emergency tasks, while elite athletes with higher HRV scores tend to show greater mental resilience during competition.[10,11] The same relationship has been observed in eSports, where elite gamers have peak HRs exceeding 160 bpm during tournament play, reflecting a sympathetic stress state similar to that of Formula 1 drivers.[12-14] These stress-induced elevations in HR are often accompanied by increased cortisol levels, supporting the association between cognitive demand and physiological stress.[13,15]

Chess is a compelling context for studying psychophysiological performance under stress. Despite its lack of physical exertion, tournament chess has been shown to elicit marked autonomic changes.[4,16] Troubat *et al.* conducted a physiological study on competitive chess, finding that players experienced an average HR increase from 75 to 86 bpm during gameplay, along with a rise in systolic blood pressure and a spike in respiratory exchange ratio (RER), suggesting an acute metabolic stress response. These changes occurred despite players being physically still, emphasizing the role of cognitive demand in activating the sympathetic nervous system.[4]

Subsequent research has expanded on these findings, using HRV and EEG to record real-time psychophysiological responses in chess players during problem-solving.[16-19] Villafaina *et al.* reported that HRV declined significantly as players were exposed to increasingly complex chess problems. Interestingly, higher-rated players were able to maintain higher HRV values than their lower-rated counterparts even under intense conditions, indicating better autonomic regulation.[19] This aligns with more general findings in performance psychology that suggest expert performers are not only more skilled but also better able to manage physiological arousal.[20]

Pereira *et al.* extended this line of research to younger populations, showing that adolescent chess players exhibited both

DOI: 10.36838/v8i2.81

decreased HRV and increased EEG theta power during time-constrained decision-making, consistent with heightened cognitive load and stress.[17] These findings suggest that chess performance under pressure involves a coordinated response between central (brain) and peripheral (autonomic) systems and that physiological markers can serve as valid indicators of cognitive strain.[21]

In chess, it has been found that time pressure (operationally defined as the phase of the game when the player's remaining time on the clock is less than 30 seconds) increases stress responses. Blitz chess, which limits players to 3-10 minutes per side, is known to put players under significant HR acceleration and neural activation.[18,22] Amidzic *et al*. showed that rapid chess gameplay elicited increased gamma-band EEG bursts in frontal and temporal regions, reflecting enhanced pattern retrieval and working memory load.[23] More recently, studies have shown that during 1-minute games, players exhibit elevated theta activity in parietal and occipital regions, as well as increased right-hemisphere activation associated with visuospatial processing.[18] These neural changes suggest that extreme time constraints force players to rely more heavily on intuition and pattern recognition, processes that may be vulnerable to stress overload.[24,25]

Although the relationship between stress and performance is well-supported, key topics remain underinvestigated.[26] Specifically, while prior studies have identified general correlations between HR/HRV and chess performance, few have established a quantitative threshold at which HR reliably predicts cognitive decline.[19] Unlike physical sports, where performance is often known to degrade above certain HR levels (e.g., >160 bpm in shooting sports), there is no established physiological "cutoff" in cognitive games like chess.[27] Moreover, existing literature has not adequately addressed whether age moderates the stress–performance relationship. Given that aging is associated with decreased HRV and increased cardiovascular stiffness, older players may be more susceptible to performance degradation at elevated HRs.[6,28]

The present study addresses these gaps by evaluating whether elevated heart rate impairs move-by-move decision accuracy in elite chess players and whether age influences this relationship. We hypothesized that HRs exceeding 130 bpm would be associated with a statistically significant drop in accuracy, particularly under time pressure.[6] We also explored whether older players exhibit heightened HR reactivity and reduced performance compared to younger players.[6,28] Using a dataset of 50 elite-level chess games with publicly available heart rate and move accuracy data, we analyzed how physiological stress correlated with decision-making performance. By combining game accuracy analysis with physiological data, this study contributes novel quantitative evidence on how elevated arousal influences decision quality in cognitively demanding environments like chess.

■ Methods

Participants:

This study analyzed 50 games played by 50 chess grandmasters, all with FIDE Elo ratings (official rating of the International Chess Federation) above 2500 at the time of competition (mean Elo rating: 2653 ± 37 Elo (SD)).[29] All players held the title of Grandmaster (GM). Games were sourced from publicly available YouTube broadcasts of elite chess tournaments between 2022 and 2024, in which real-time biometric heart rate data was visible on-screen. All games were played under blitz time controls, defined as 3 to 10 minutes per side.[22] For anonymity, no player names or identifying details were recorded. Player ages ranged from 18 to 55 years (mean age: 27.6 ± 6.9 years (SD)) and were grouped into younger (≤25 years) and older (>25 years) categories based on previous research examining age-related variability in chess performance and stress responses.[30] This division also approximately reflected the median of our sample. Of the 50 grandmasters, 22 were aged 25 or younger, while 28 were older than 25.

Data Acquisition and Preprocessing:

While several prior studies have used heart rate variability (HRV) as a marker of cognitive load and stress regulation, the present study focuses exclusively on real-time heart rate (HR) due to the nature of the publicly available data.[6] Heart rate (HR) data was collected every three moves for each game from the visible biometric data displayed during the public YouTube broadcasts. HR data was measured by wrist sensors worn by the players and displayed real-time heart rate values in beats per minute (bpm). HR values were recorded and matched to corresponding moves by cross-referencing player clocks visible on-screen and the move sequence as tracked in real-time using Chess.com's live game interface.[31] HR values were recorded both for the entire game and separately for moves made during time pressure (≤30 seconds remaining on the clock) and during normal play (>30 seconds remaining on the clock).

Game accuracy was calculated using Chess.com's accuracy scoring system, CAPS2 (Computer Accuracy Precision Score), which evaluates every move using the Stockfish engine, one of the strongest chess engines in the world, and compares it to the best available option in the position.[32,33] The official accuracy metric considers how close each move is to the engine's top choice, producing a game-level percentage score ranging from 0 to 100. This methodology has been widely adopted in online chess analysis.[33] Accuracy values were recorded both for the entire game and separately for moves made during time pressure (≤30 seconds remaining on the clock) and during normal play (>30 seconds remaining on the clock).

Time pressure was operationally defined as any phase of the game in which the player's remaining time on the clock dropped below 30 seconds. This was chosen to reflect a widely accepted threshold in competitive chess, under which cognitive performance may be compromised due to rapid decision-making and limited working memory.[34] Both HR and move accuracy were independently measured during these time-constrained periods.

To define heart rate groups, we categorized players into Low-to-Moderate HR (≤130 bpm) and High HR (>130 bpm) categories based on their average game heart rate. This threshold was chosen based on evidence from the psychophysiological literature indicating that heart rates above 130 bpm are associated with a shift toward sympathetic dominance and decreased cognitive efficiency under stress.[2,35] This division also approximately reflected the median of our sample.

Player ages were sourced from FIDE.com and other tournament coverage, allowing players to be categorized as younger (≤25 years) or older (>25 years).[36] All extracted data was compiled into a dataset that included the following variables: mean HR per game, mean HR during time pressure (if applicable), mean HR during normal play (bpm), total game accuracy (%), accuracy during time pressure (if applicable), accuracy during normal play (%), player age group, and HR group.

Statistical Analysis:

All statistical analyses were conducted using Python and associated scientific libraries. A significance threshold of $p < 0.05$ was used for all tests, and 95% confidence intervals (CI) were calculated where applicable. Values are reported as means ± standard error (SE) unless otherwise noted.

To compare decision-making accuracy between players with elevated heart rates and those with lower heart rates, Welch's two-sample t-test was used. This test was selected due to potential unequal variances between groups and was applied to compare mean game accuracy between the High HR group (>130 bpm) and the Low-to-Moderate HR group (≤130 bpm). 2 paired t-tests were conducted to analyze within-player changes in both heart rate and accuracy during time pressure. Specifically, players' mean HR and accuracy values during normal play (>30 seconds remaining) were compared to values recorded during time pressure periods (≤30 seconds remaining) to evaluate physiological and performance changes under stress.

To investigate the relationship between heart rate values and decision-making accuracy across the dataset, a Pearson correlation analysis was performed. In addition, a simple linear regression model was calculated to find whether heart rate could significantly predict accuracy using the equation:

$$\text{Accuracy} = \beta_0 + \beta_1 \cdot HR + \varepsilon$$

Age-based differences in heart rate and accuracy were analyzed using independent two-sample t-tests, comparing outcomes between players aged ≤25 years and those >25 years.

■ Result and Discussion

Results:

Descriptive Statistics:

The overall mean heart rate across 50 games was 129.6 bpm (SE = 2.1), and the average move accuracy was 91.2% (SE = 1.2). Mean heart rate increased from 123.4 bpm (SE = 2.2) during normal play to 137.6 bpm (SE = 2.1) under time pressure (≤30 seconds). Similarly, move accuracy dropped from 94.1% (SE = 1.1) during normal play to 85.2% (SE = 1.4) under time pressure. Age-based differences were also observed: older players (>25 years) had a higher mean heart rate 132.9 bpm (SE = 2.3), and lower accuracy, 90.6% (SE = 1.6), than younger players (≤25 years), who averaged 123.8 bpm (SE = 2.6) and 91.7% (SE = 1.5), respectively.

Result 1: Accuracy by Heart Rate Group:

A Welch's two-sample t-test was performed to compare mean move accuracy between players with high average heart rates (>130 bpm) and those with lower heart rates (≤130 bpm). The High HR group (n = 26) had a significantly lower accuracy (Mean = 87.9%, SE = 1.5) compared to the Low HR group (n = 24) (Mean = 93.2%, SE = 1.2), t = 2.76, p = 0.0082 (Figure 1), indicating that elevated heart rate is associated with a reduction in decision-making performance.

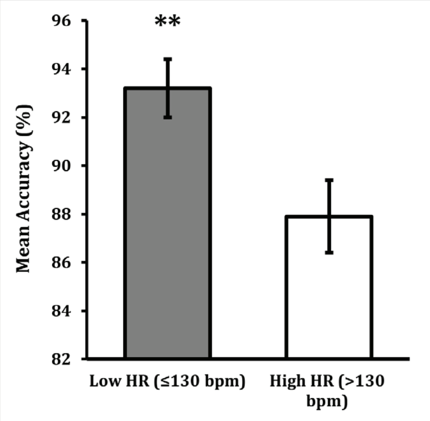


Figure 1: Elevated heart rate is associated with reduced move accuracy in elite chess players. Bar graph showing mean ± SE move accuracy (%) for players with low (≤130 bpm) and high (>130 bpm) average heart rates during tournament games (n = 50 games total). Mean accuracy was significantly lower in the high HR group (87.9 ± 1.5%) compared to the low HR group (93.2 ± 1.2%). Welch's two-sample t-test, **p < 0.01.

Result 2: Accuracy Under Time Pressure vs. Normal Play:

A paired t-test was conducted to compare player accuracy during normal play (>30 seconds on the clock) and time pressure intervals (≤30 seconds on the clock) (n = 50 paired observations). Accuracy was significantly lower during time pressure (Mean = 85.2%, SE = 1.4) compared to normal conditions (Mean = 94.1%, SE = 1.1), t = 4.99, p < 0.001 (Figure 2), indicating that high time pressure significantly impairs move quality.

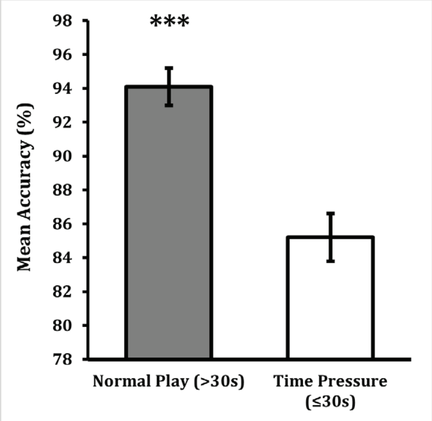


Figure 2: Time pressure significantly impairs move accuracy in chess games. Bar graph showing mean ± SE move accuracy (%) during normal play (>30 seconds) versus time pressure intervals (≤30 seconds remaining) during tournament games (n = 50 paired observations). Accuracy was significantly lower under time pressure (85.2 ± 1.4%) compared to normal conditions (94.1 ± 1.1%). Paired t-test, ***p < 0.001.

DOI: 10.36838/v8i2.81

Result 3: Heart Rate Under Time Pressure vs. Normal Play:

A paired t-test comparing heart rate in normal play versus time pressure revealed a significant increase under time pressure conditions (n = 50 paired observations). Mean HR rose from 123.4 bpm (SE = 2.2) during normal play to 137.6 bpm (SE = 2.1) during time pressure, t = 4.67, p < 0.001 (Figure 3), confirming that time pressure elicits physiological stress responses in elite players.

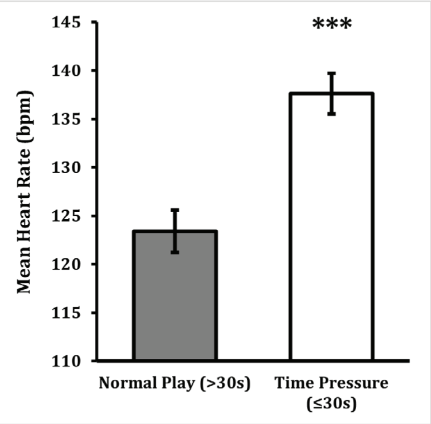


Figure 3: Time pressure significantly increases heart rate in elite chess players. Bar graph showing mean ± SE heart rate (bpm) during normal play (>30 seconds) versus time pressure intervals (≤30 seconds remaining) during tournament games (n = 50 paired observations). Mean heart rate increased significantly from 123.4 ± 2.2 bpm during normal play to 137.6 ± 2.1 bpm during time pressure. Paired t-test, ***p < 0.001.

Result 4: Correlation Between Heart Rate and Accuracy:

A Pearson correlation analysis showed a significant negative relationship between heart rate and accuracy across all games (n = 50), r = –0.565, p < 0.001. Linear regression further confirmed this association, with heart rate significantly predicting move accuracy (β = –0.3513, p < 0.001) (Figure 4). The regression model explained 31.9% of the variance in accuracy (R^2 = 0.319), with the following equation:

$$Accuracy = -0.3513 \times HR + 136.42$$

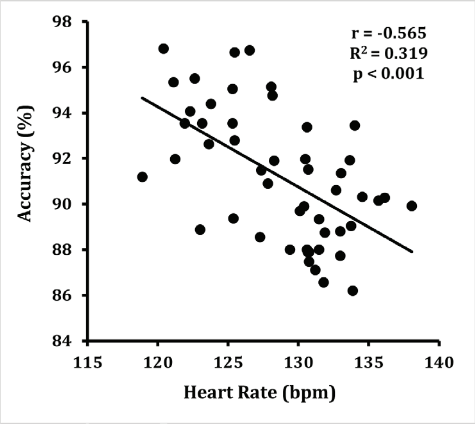


Figure 4: Heart rate is negatively associated with move accuracy across games. Scatter plot with linear regression line showing the relationship between heart rate (bpm) and move accuracy (%) across all tournament games (n = 50). A Pearson correlation analysis revealed a significant negative association between heart rate and accuracy (r = –0.565, p < 0.001). Linear regression confirmed that heart rate significantly predicted move accuracy (β = –0.3513, p < 0.001), explaining 31.9% of the variance in accuracy (R^2 = 0.319). The regression equation was Accuracy = -0.3513 x HR + 136.42.

Result 5: Age Group Differences in HR and Accuracy:

Independent T-tests were conducted to examine age-related differences in HR and accuracy. Older players (>25 years) (n=28) showed significantly higher mean heart rates (Mean = 132.9 bpm, SE = 2.3) compared to younger players (≤25 years) (n=22) (Mean = 123.8 bpm, SE = 2.6), t = 2.62, p = 0.0119 (Figure 5A). Additionally, while older players had lower accuracy (Mean = 90.6%, SE = 1.6) than younger players (Mean = 91.7%, SE = 1.5), t = 0.5, p = 0.62 (Figure 5B), this difference was not statistically significant.

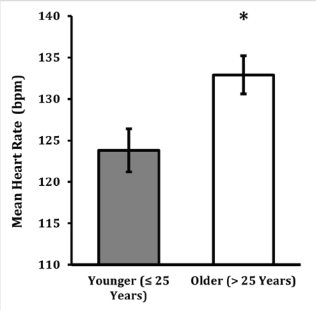


Figure 5A: Heart rate by age group. Bar graph showing mean ± SE heart rate (bpm) for younger (≤25 years) and older (>25 years) chess players (n = 50). Older players exhibited significantly higher heart rates (132.9 ± 2.3 bpm) compared to younger players (123.8 ± 2.6 bpm). Independent t-test, *p < 0.05.

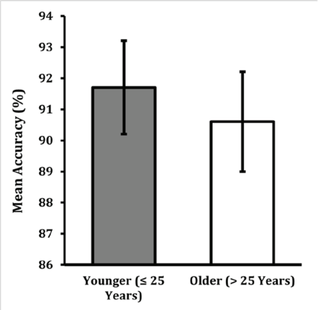


Figure 5B: Move accuracy by age group. Bar graph showing mean ± SE move accuracy (%) for younger (≤25 years) and older (>25 years) chess players (n = 50). Accuracy was slightly lower in older players (90.6 ± 1.6%) than in younger players (91.7 ± 1.5%), but this difference was not statistically significant. Independent t-test, ns.

Discussion:

This study provides quantitative evidence that elevated physiological arousal, measured via heart rate (HR), is significantly associated with reduced decision-making accuracy in elite chess players. The hypothesis that sustained HRs above a threshold of 130 bpm would correspond to lower accuracy was supported across multiple statistical analyses. Players with higher average HRs performed significantly worse than their lower HR counterparts, both in overall game accuracy and under time-pressure conditions. These findings align with established psychophysiological theories suggesting that excessive sympathetic activation impairs cognitive functioning and working memory, particularly under acute stress and cognitive load.[1,37,38]

Time pressure emerged as a particularly important stressor. In blitz-format games, players had limited time to calculate complex positions, and their physiological data confirmed a significant increase in HR during these moments. This autonomic response was accompanied by a significant drop in move accuracy, supporting the idea that stress reduces concentration, encourages faster but less thoughtful decisions, and makes deep analysis harder.[39,40]

The relationship between heart rate and performance was further supported by a significant negative correlation and a significant predictive regression model. Unlike binary threshold models that assume a fixed cutoff, our linear model showed that decision accuracy declines continuously as HR increases. These results mirror the broader literature on cognitive fatigue and physiological dysregulation, and they point toward HR as a real-time biomarker for cognitive strain.[1,41-43] Monitoring HR could thus serve as an indicator of internal cognitive load, especially in fast-paced decision-making contexts.[41,42]

Age-based differences further supported this relationship. Older players (>25 years) exhibited significantly higher HRs and slightly lower move accuracy (not statistically significant) compared to younger counterparts, suggesting they may be more physiologically reactive to stress or experience a sharper decline in performance when under pressure. This aligns with literature linking age to decreased heart rate variability (HRV), reduced prefrontal efficiency, and diminished autonomic recovery.[6,28] The implication is that even among elite competitors, biological age may moderate how well individuals handle cognitive stress.[28] This highlights the potential benefit of individualized stress-monitoring tools in cognitive competition.[44]

The implications of this research extend beyond chess. While the game offers a model for studying cognition under pressure, the relationships researched here are relevant to other high-stakes mental settings. Fields such as eSports, air traffic control, and high-pressure workplaces rely on rapid, accurate judgments under time constraints.[45-47] Training protocols that incorporate biofeedback, stress-regulation techniques, or HR monitoring tools could help individuals maintain cognitive efficiency under stress.[48]

Although the findings are statistically significant, there are some limitations to consider. Firstly, HR data was extracted from publicly available tournament broadcasts and lacked continuous tracking of HR. Secondly, while Chess.com's accuracy metric is widely used, it simplifies decision quality to engine-based comparisons and may not capture psychological or strategic nuances of specific positions, and it does not account for factors such as position difficulty, transparency, or misleading cases. Thirdly, this was a cross-sectional, observational study. While HR and accuracy were correlated, causality cannot be inferred. Finally, the sample consisted exclusively of grandmasters. Hence, the findings may not be generalizable to the broader chess community.

Conclusion

In summary, this study provides strong empirical evidence that elevated heart rate, a marker for physiological stress, is significantly associated with reduced decision-making accuracy in elite-level chess. By analyzing biometric and game performance data across 50 games, we confirmed that players with higher heart rates (>130 bpm) consistently played with lower accuracy, particularly under time pressure.

Regression and correlation analyses further established an inverse relationship between HR and cognitive precision, suggesting that even non-physical mental performance is sensitive to autonomic arousal.

Age was also a moderating factor, as older players exhibited significantly higher stress reactivity and slightly lower accuracy, highlighting individual physiological variability in high-pressure decision-making. These findings improve our understanding of performance under cognitive pressure and have broader implications for other domains, where rapid and accurate decisions are critical.[47]

Future research should build upon these findings by integrating continuous heart rate monitoring, heart rate variability (HRV), and neurocognitive markers such as EEG to better understand real-time stress responses.[20,49] Expanding the dataset to include players of varying skill levels, gender, and tournament formats will help increase the generalizability of these results.

Acknowledgments

We would like to express our deepest gratitude to the contributors of biometric chess data, as well as the developers of Chess.com, for their live-game interface and accuracy tools.

References

1. Arnsten, A. F. T. Stress signalling pathways that impair prefrontal cortex structure and function. *Nature Reviews. Neuroscience* 2009, *10* (6), 410–422. https://doi.org/10.1038/nrn2648.
2. Diamond, D. M.; Campbell, A. M.; Park, C. R.; Halonen, J.; Zoladz, P. R. The Temporal Dynamics model of emotional memory processing: a synthesis on the neurobiological basis of Stress-Induced amnesia, flashbulb and traumatic memories, and the Yerkes-Dodson Law. *Neural Plasticity* 2007, *2007*, 1–33. https://doi.org/10.1155/2007/60803.
3. Hermans, E. J.; Henckens, M. J. A. G.; Joëls, M.; Fernández, G. Dynamic adaptation of large-scale brain networks in response to acute stressors. *Trends in Neurosciences* 2014, *37* (6), 304–314. https://doi.org/10.1016/j.tins.2014.03.006.
4. Troubat, N.; Fargeas-Gluck, M.-A.; Tulppo, M.; Dugué, B. The stress of chess players as a model to study the effects of psychological stimuli on physiological responses: an example of substrate oxidation and heart rate variability in man. *European Journal of Applied Physiology* 2008, *105* (3), 343–349. https://doi.org/10.1007/s00421-008-0908-2.
5. Lehrer, P. M.; Gevirtz, R. Heart rate variability biofeedback: how and why does it work? *Frontiers in Psychology* 2014, *5*. https://doi.org/10.3389/fpsyg.2014.00756.
6. Thayer, J. F.; Åhs, F.; Fredrikson, M.; Sollers, J. J.; Wager, T. D. A meta-analysis of heart rate variability and neuroimaging studies: Implications for heart rate variability as a marker of stress and health. *Neuroscience & Biobehavioral Reviews* 2011, *36* (2), 747–756. https://doi.org/10.1016/j.neubiorev.2011.11.009.
7. Billman, G. E. Heart rate variability ? A historical perspective. *Frontiers in Physiology* 2011, *2*. https://doi.org/10.3389/fphys.2011.00086.
8. Smeets, T. Autonomic and hypothalamic–pituitary–adrenal stress resilience: Impact of cardiac vagal tone. *Biological Psychology* 2010, *84* (2), 290–295. https://doi.org/10.1016/j.biopsycho.2010.02.015.
9. Forte, G.; Morelli, M.; Grässler, B.; Casagrande, M. Decision making and heart rate variability: A systematic review. *Applied Cognitive Psychology* 2021, *36* (1), 100–110. https://doi.org/10.1002/acp.3901.
10. Corrigan, S. L.; Roberts, S.; Warmington, S.; Drain, J.; Main, L. C. Monitoring stress and allostatic load in first responders and tactical operators using heart rate variability: a systematic review. *BMC Public Health* 2021, *21* (1). https://doi.org/10.1186/s12889-021-11595-x.
11. Overton, F. *HRV: The Endurance Athlete's Complete Guide*. FasCat Coaching. https://fascatcoaching.com/blogs/training-tips/hrv-heart-rate-variability?

12. Krell, J. *Is esports a sport? Researchers undecided*. Global Sport Matters. https://globalsportmatters.com/culture/2019/09/18/is-esports-a-sport-researchers-undecided/?utm_.

13. Zimmer, R. T.; Haupt, S.; Heidenreich, H.; Schmidt, W. F. J. Acute effects of esports on the cardiovascular system and energy expenditure in amateur esports players. *Frontiers in Sports and Active Living* 2022, *4*. https://doi.org/10.3389/fspor.2022.824006.

14. Tornaghi, M.; Vandoni, M.; Zaccaria, D.; D'Antona, G.; Codella, R.; Lovecchio, N. Heart rate profiling in Formula 1 race: A real-time case. *Science & Sports* 2023, *38* (7), 736–740. https://doi.org/10.1016/j.scispo.2022.10.005.

15. Dickerson, S. S.; Kemeny, M. E. Acute Stressors and Cortisol Responses: A Theoretical Integration and Synthesis of Laboratory Research. *Psychol. Bull*. 2004, *130* (3), 355–391. https://doi.org/10.1037/0033-2909.130.3.355.

16. Fuentes-García, J. P.; Villafaina, S.; Collado-Mateo, D.; De La Vega, R.; Olivares, P. R.; Clemente-Suárez, V. J. Differences Between High vs. Low Performance Chess Players in Heart Rate Variability During Chess Problems. *Frontiers in Psychology* 2019, *10*. https://doi.org/10.3389/fpsyg.2019.00409.

17. Fuentes-García, J. P.; Pereira, T.; Castro, M. A.; Santos, A. C.; Villafaina, S. Psychophysiological stress response of adolescent chess players during problem-solving tasks. *Physiology & Behavior* 2019, *209*, 112609. https://doi.org/10.1016/j.physbeh.2019.112609.

18. Villafaina, S.; Collado-Mateo, D.; Cano-Plasencia, R.; Gusi, N.; Fuentes, J. P. Electroencephalographic response of chess players in decision-making processes under time pressure. *Physiology & Behavior* 2018, *198*, 140–143. https://doi.org/10.1016/j.physbeh.2018.10.017.

19. Villafaina, S.; Castro, M. A.; Pereira, T.; Santos, A. C.; Fuentes-García, J. P. Neurophysiological and autonomic responses of high and low level chess players during difficult and easy chess endgames – A quantitative EEG and HRV study. *Physiology & Behavior* 2021, *237*, 113454. https://doi.org/10.1016/j.physbeh.2021.113454.

20. Ericsson, K. A. (2006). The influence of experience and deliberate practice on the development of superior expert performance. In *The Cambridge Handbook of Expertise and Expert Performance* (pp. 683–706). Cambridge University Press.

21. Critchley, H. D. Neural mechanisms of autonomic, affective, and cognitive integration. *The Journal of Comparative Neurology* 2005, *493* (1), 154–166. https://doi.org/10.1002/cne.20749.

22. *Fédération Internationale des Échecs (FIDE). FIDE Handbook – Laws of Chess: Appendix B. Blitz*. https://handbook.fide.com/chapter/B02RBRegulations2024.

23. Amidzic, O.; Riehle, H. J.; Fehr, T.; Wienbruch, C.; Elbert, T. Pattern of focal γ-bursts in chess players. *Nature* 2001, *412* (6847), 603. https://doi.org/10.1038/35088119.

24. Guntz, T.; Crowley, J.; Vaufreydaz, D.; Balzarini, R.; Dessus, P. *The Role of Emotion in Problem Solving: First Results from Observing Chess*. arXiv.org. https://arxiv.org/abs/1810.11094.

25.Klein,G.SourcesofPower:Howpeoplemakedecisions.*ResearchGate* 2001. https://doi.org/10.1061/(ASCE)1532-6748(2001)1:1(21).

26. Chen, B.; Wang, L.; Li, B.; Liu, W. Work stress, mental health, and employee performance. *Frontiers in Psychology* 2022, *13*. https://doi.org/10.3389/fpsyg.2022.1006580.

27. Açıkada, C.; Hazır, T.; Asçı, A.; Aytar, S. H.; Tınazcı, C. Effect of heart rate on shooting performance in elite archers. *Heliyon* 2019, *5* (3), e01428. https://doi.org/10.1016/j.heliyon.2019.e01428.

28. Lin, F.; Heffner, K.; Mapstone, M.; Chen, D.-G.; Porsteisson, A. Frequency of mentally stimulating activities modifies the relationship between cardiovascular reactivity and executive function in old age. *American Journal of Geriatric Psychiatry* 2013, *22* (11), 1210–1221. https://doi.org/10.1016/j.jagp.2013.04.002.

29. Fédération Internationale des Échecs (FIDE). *FIDE Handbook – Rating Regulations*. https://handbook.fide.com/chapter/B022024.

30. Grabner, R. H.; Stern, E.; Neubauer, A. C. Individual differences in chess expertise: A psychometric investigation. *Acta Psychologica* 2006, *124* (3), 398–420. https://doi.org/10.1016/j.actpsy.2006.07.008.

31. Chess.com. *Live Chess Platform and Game Archive*. https://www.chess.com/watch

32. Stockfish Chess Engine. *Stockfish – Open Source Chess Engine*. https://stockfishchess.org

33. *How is accuracy in Analysis determined? | Chess.com Help Center*. https://support.chess.com/en/articles/8708970-how-is-accuracy-in-analysis-determined.

34. Van Harreveld, F.; Wagenmakers, E.-J.; Van Der Maas, H. L. J. The effects of time pressure on chess skill: an investigation into fast and slow processes underlying expert performance. *Psychological Research* 2006, *71* (5), 591–597. https://doi.org/10.1007/s00426-006-0076-0.

35. Glier, S.; Campbell, A.; Corr, R.; Pelletier-Baldelli, A.; Yefimov, M.; Guerra, C.; Scott, K.; Murphy, L.; Bizzell, J.; Belger, A. Coordination of autonomic and endocrine stress responses to the Trier Social Stress Test in adolescence. *Psychophysiology* 2022, *59* (9). https://doi.org/10.1111/psyp.14056.

36. Fédération Internationale des Échecs (FIDE). *Player Database and Ratings*. https://ratings.fide.com

37. Qin, S.; Cousijn, H.; Rijpkema, M.; Luo, J.; Franke, B.; Hermans, E. J.; Fernández, G. The effect of moderate acute psychological stress on working memory-related neural activity is modulated by a genetic variation in catecholaminergic function in humans. *Frontiers in Integrative Neuroscience* 2012, *6*. https://doi.org/10.3389/fnint.2012.00016.

38. Schoofs, D.; Wolf, O. T.; Smeets, T. Cold pressor stress impairs performance on working memory tasks requiring executive functions in healthy young men. *Behavioral Neuroscience* 2009, *123* (5), 1066–1075. https://doi.org/10.1037/a0016980.

39. Porcelli, A. J.; Delgado, M. R. Stress and decision making: effects on valuation, learning, and risk-taking. *Current Opinion in Behavioral Sciences* 2016, *14*, 33–39. https://doi.org/10.1016/j.cobeha.2016.11.015.

40. Pabst, S.; Brand, M.; Wolf, O. T. Stress and decision making: A few minutes make all the difference. *Behavioural Brain Research* 2013, *250*, 39–45. https://doi.org/10.1016/j.bbr.2013.04.046.

41. Goodman, S. P. J.; Collins, B.; Shorter, K.; Moreland, A. T.; Papic, C.; Hamlin, A. S.; Kassman, B.; Marino, F. E. Approaches to inducing mental fatigue: A systematic review and meta-analysis of (neuro)physiologic indices. *Behavior Research Methods* 2025, *57* (4). https://doi.org/10.3758/s13428-025-02620-7.

42. Arutyunova, K. R.; Bakhchina, A. V.; Konovalov, D. I.; Margaryan, M.; Filimonov, A. V.; Shishalov, I. S. Heart rate dynamics for cognitive load estimation in a driving simulation task. *Scientific Reports* 2024, *14* (1). https://doi.org/10.1038/s41598-024-79728-x.

43. Boksem, M. A. S.; Meijman, T. F.; Lorist, M. M. Effects of mental fatigue on attention: An ERP study. *Cognitive Brain Research* 2005, *25* (1), 107–116. https://doi.org/10.1016/j.cogbrainres.2005.04.011.

44. Bolpagni, M.; Pardini, S.; Dianti, M.; Gabrielli, S. Personalized Stress Detection Using Biosignals from Wearables: A Scoping Review. *Sensors* 2024, *24* (10), 3221. https://doi.org/10.3390/s24103221.

45. Berga, D.; Pereda, A.; Eleonora, D. F.; Nandi, A.; Febrer, E.; Reverte, M.; Russo, L. *Measuring arousal and stress physiology on Esports, a League of Legends case study*. arXiv.org. https://arxiv.org/abs/2302.14269.

46. Li, W.; Zhang, J.; Kearney, P. Psychophysiological coherence training to moderate air traffic controllers' fatigue on rotating roster. *Risk Analysis* 2022, *43* (2), 391–404. https://doi.org/10.1111/risa.13899.
47. Prinsloo, G. E.; Rauch, H. G. L.; Lambert, M. I.; Muench, F.; Noakes, T. D.; Derman, W. E. The effect of short duration heart rate variability (HRV) biofeedback on cognitive performance during laboratory induced cognitive stress. *Applied Cognitive Psychology* 2010, *25* (5), 792–801. https://doi.org/10.1002/acp.1750.
48. Iodice, P.; Cannito, L.; Chaigneau, A.; Palumbo, R. Learned self-regulation in top-level managers through neurobiofeedback training improves decision making under stress. *Scientific Reports* 2022, *12* (1). https://doi.org/10.1038/s41598-022-10142-x.
49. Forte, G.; Casagrande, M. The intricate brain–heart connection: The relationship between heart rate variability and cognitive functioning. *Neuroscience* 2024. https://doi.org/10.1016/j.neuroscience.2024.12.004.

■ Authors

Yash Jayesh Laddha is a senior at Greenwood High International School. He is an international chess player (Candidate Master) who is passionate about biological research, particularly in molecular biology and stress physiology. He is fascinated by the intersection of AI and biology and hopes to pursue it in the future.

Shubh Jayesh Laddha is a senior at Delhi Public School East, Bangalore. He is an international chess player (FIDE Master) who is passionate about data science, particularly in AI and ML research. He hopes to pursue mathematics integrated with AI and ML. In his free time, he enjoys playing squash.

IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Modeling Astrocyte Influence on Risperidone Response in Neuropsychiatric Treatment - Using COPASI

Aishwarya Ashok

Central Bucks High School South, 1100 Folly Rd, Warrington, PA, 18976, USA; aishwaryaashok1414@gmail.com

ABSTRACT: Bipolar disorder and schizophrenia are prevalent neuropsychiatric disorders. Risperidone is an antipsychotic drug that has variability among patient responses to it. Astrocytes play a critical role in regulating neurotransmitter levels, particularly glutamate. Glutamate is the primary excitatory neurotransmitter; elevated glutamate concentrations in the synapse can lead to excitotoxicity, potentially influencing the efficacy of risperidone. This research aims to find how astrocytic glutamate uptake influences risperidone's efficacy. A computational simulation was developed using COPASI, a biochemical pathway modeling tool, to explore the dynamics of dopamine and risperidone binding to the D2 under a high glutamate concentration of 5 µmol/L. Two trials were conducted: a normal astrocyte model with active glutamate uptake, and a defective astrocyte model, where glutamate uptake was absent (V_{max} set to 0). Other values were held constant. The results showed that despite simulating excitotoxic conditions with high glutamate levels in both models, there was no significant change in risperidone's binding efficacy to the D2 receptor compared to the normal astrocyte model. These observations suggest that, under the utilized experimental conditions, glutamate uptake by astrocytes does not directly influence the rate of risperidone binding to the D2 receptor. This suggests that there may be other mechanisms influencing risperidone response.

KEYWORDS: Behavioral and Social Sciences, Neuroscience, Cellular and Molecular Biology, Neurobiology, Computational Biology and Bioinformatics, Computational Pharmacology.

Introduction

Schizophrenia and bipolar disorder are two of the most prevalent and severe neuropsychiatric disorders, each characterized by significant disruptions in mood, cognition, and behavior. Schizophrenia is expressed through symptoms such as hallucinations, delusions, disorganized thinking, and cognitive deficits, while bipolar disorder involves episodes of mania and/or depression, often accompanied by impairments in executive function, emotional regulation, and memory.[1] Despite their differences, both disorders share certain neurochemical variances that contribute to the dysregulation in dopamine and glutamate signaling.

Risperidone, a commonly used atypical antipsychotic drug, is prescribed to ease the symptoms of neuropsychiatric conditions like schizophrenia and bipolar disorder. Its primary mechanism of action involves antagonism of dopamine D2 receptors (D2Rs), reducing excess dopamine activity in key brain regions such as the striatum and prefrontal cortex.[2] However, patient response and outcomes with risperidone vary significantly between individuals. Some patients experience effective symptom relief, while others suffer from severe side effects such as cognitive impairment, metabolic dysfunction, or psychomotor agitation. In some cases, risperidone fails to produce a meaningful therapeutic response at all.[3] The variability in treatment response suggests that additional neurochemical factors play a role in risperidone's neurotransmitter interactions.

Recent evidence suggests that astrocytes, a specialized type of glial cell, may play a crucial but unrecognized role in maintaining antipsychotic drug efficacy. Traditionally thought to be passive support cells, astrocytes are now recognized as active regulators of neurotransmission, maintaining homeostasis within the central nervous system (CNS).[4] One of their most critical functions is the regulation of glutamate, the brain's primary excitatory neurotransmitter. Astrocytes accomplish this by expressing excitatory amino acid transporters (EAATs), which actively remove excess glutamate from the synapse and recycle it for reuse (Figure 1).[5] This process is vital for preventing glutamate excitotoxicity, a condition that results in overactivation of glutamate receptors and eventual neuronal damage or death.[6]

Maintaining homeostasis of glutamate levels is particularly relevant to behavioral health because glutamatergic dysregulation has been related to schizophrenia and bipolar disorder.[7] Excessive glutamate levels have been associated with cortical thinning, synaptic destabilization, and disruptions in neural connectivity, all of which contribute to cognitive and emotional impairments observed in these disorders.[8] Astrocytic dysfunction, whether due to reduced rates of glutamate uptake or imbalanced transporter activity, may lead to an accumulation of extracellular glutamate, which in turn can interfere with drug-to-neurotransmitter interactions and alter the efficacy of antipsychotic drugs.[9]

DOI: 10.36838/v8i2.88

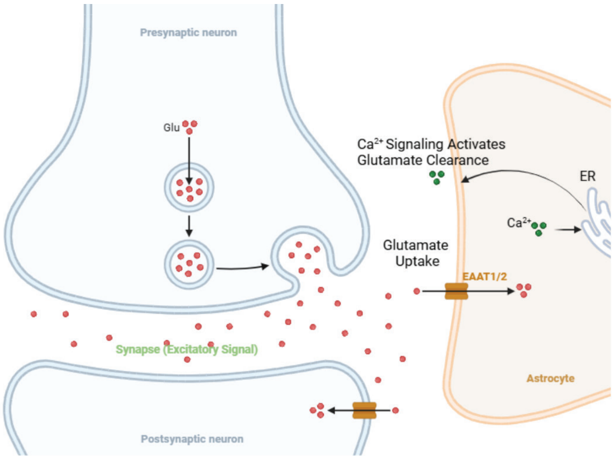


Figure 1: Model depicts Calcium Signaling and Glutamate Uptake between Neurons and Astrocyte cells. This also shows activation of Glutamate Clearance through Calcium Signaling. (Student Produced)

Glutamate, the primary excitatory neurotransmitter, is transported to neurons via the synapses. During this process, astrocytes initiate glutamate uptake to maintain stable levels of glutamate being transported. Excess levels of glutamate in the synapse may lead to excitotoxicity, resulting in neuronal damage or cell death.

The relationship between astrocytic glutamate regulation and risperidone efficacy is not fully understood, yet it may hold the key to explaining why treatment outcomes vary among individuals. If astrocytic glutamate uptake is compromised, excess extracellular glutamate may alter dopamine receptor availability and efficacy. This disruption could lead to lower risperidone binding affinity for D2Rs, thereby reducing its intended effects on dopamine regulation. Conversely, enhancing astrocytic glutamate uptake may improve drug efficacy by creating neurotransmitter balance, possibly reducing the need for higher doses and minimizing adverse symptoms of risperidone use.[10]

This study aims to investigate the role of astrocytes in influencing risperidone's pharmacological effects through glutamate regulation. By utilizing computational modeling, risperidone's interaction with dopamine D2 receptors will be simulated under two conditions: one representing normal astrocytic function with active glutamate uptake, and another representing a defective astrocyte where glutamate uptake is impaired (V_{max} = 0). This approach allows for an analysis of how astrocytic function influences drug binding kinetics, dopamine-glutamate interactions, and overall treatment efficacy.

Understanding how astrocytes contribute to the variability in risperidone response has great implications for behavioral health. It competes with the longstanding neuron-focused view of psychiatric disorders and introduces astrocytes as key regulators of drug activity. If astrocytes are found to play a significant role in risperidone's effectiveness, future pharmacological interventions could more efficiently utilize astrocyte-targeted treatments to enhance drug efficacy and develop more personalized therapeutic strategies for schizophrenia and bipolar disorder. By bridging the gap between dopamine and glutamate-based models of mental illness, this study contributes to a more holistic understanding of psychiatric disorders and their treatment, aiming to eventually improve overall patient responses and quality of life for affected individuals.

■ Methods

Hypothesis: If astrocyte-mediated signaling inhibits dopamine receptors, then astrocytes influence the efficacy of risperidone.

This research study was conducted in three phases: information gathering, computational model development, and simulation. The first phase involved an extensive review of existing literature to establish a foundation for modeling risperidone's interaction with astrocytic glutamate uptake. The goal was to identify relevant biochemical pathways, kinetic parameters, and reaction mechanisms necessary for developing an accurate computational model. Data sources included peer-reviewed journals, pharmacokinetic studies, and neurobiological research on dopamine-glutamate interactions, astrocytic glutamate transport, and risperidone's pharmacodynamics. Various experimental studies were used to find specific parameters, such as the binding affinities of risperidone to dopamine D2 receptors and the kinetics of glutamate uptake. These values allowed for defining the initial conditions, rate laws, and reaction equations used in the computational model.

Once the necessary data were collected, a computational model was developed to simulate the biochemical interactions between risperidone, dopamine, astrocytes, and glutamate uptake. Biochemical simulation software COPASI (v4.36) was used to depict specific drug-to-neurotransmitter interactions. The model was designed to capture the role of astrocytic glutamate clearance in maintaining a balance in neurotransmitter activity and its potential influence on risperidone's pharmacodynamics. Ligand-receptor interactions were incorporated by modeling the binding of risperidone and dopamine to dopamine D2 receptors using Michaelis-Menten irreversible kinetics, with K_{on} and K_{off} rates pulled from literature sources (Table 2). Glutamate uptake by astrocytes was modeled using Michaelis-Menten rate laws to represent EAAT-mediated glutamate clearance. Two conditions were created: one where astrocytes actively removed extracellular glutamate and another where glutamate uptake was impaired (V_{max} = 0) to simulate astrocytic dysfunction. The model also tracked changes in extracellular glutamate and dopamine levels over time, allowing for an analysis of how astrocytic function affects risperidone-D2 receptor interactions. The initial concentrations for dopamine, risperidone, D2 receptors, and glutamate were set based on past experimental data (Table 1). Each reaction was parameterized with literature-derived rate constants, K_m, and V_{max} values to ensure accuracy. After constructing the model, preliminary simulations were conducted to confirm that the system behaved as expected before experimental simulations were run.

The final phase involved running simulations in COPASI to examine the effects of astrocytic glutamate uptake on risperidone's binding efficacy to dopamine D2 receptors. The model was simulated for 600 seconds under two conditions: one representing normal astrocyte function, where glutamate uptake remained active, and another where glutamate uptake was absent (V_{max} = 0) to simulate impaired astrocyte function. Key variables, such as glutamate clearance rates, dopamine fluctuations, and risperidone receptor binding levels, were recorded for analysis. Data were exported from COPASI and visualized using Microsoft Excel to generate graphs comparing neurotransmitter levels and receptor occupancy across different astrocytic conditions. The results were then analyzed to determine whether glutamate uptake influences risperidone's ability to decrease dopamine signaling, providing insight into how astrocytic function may contribute to treatment variability in schizophrenia and bipolar disorder.

Table 1: Shows the key molecular components and initial conditions utilized in the COPASI simulation of the effects of glutamate uptake on risperidone's binding efficacy to D2 receptors.

Molecule	Initial Concentration	Role
D2-DA Complex	0 M	DA-bound receptor
D2 Receptor (D2)	1.0 e^{-8} M	Dopamine Receptor
Risperidone (RISP)	2.51e^{-9} M	Antipsychotic drug
D2-RISP Complex	0 M	RISP-bound receptor
Glutamate (Glu_ext)	5.0e^{-8}M	Excitatory neurotransmitter

Table 2: Shows the kinetic values utilized to replicate the movement during the various biochemical interactions. The biochemical interactions modeled include dopamine to D2 receptor binding, risperidone to D2 receptor binding in competition with dopamine, and astrocytic glutamate uptake.

Reaction	K_{on}	K_{off}	V_{max}	K_m
DA + D2 →D2-DA	2.75 x 10^4 $M^{-1}s^{-1}$	0.197 s^{-1}	–	–
RISP + D2 → D2-RISP	1.65 x 10^6 $M^{-1}s^{-1}$	1.86 x 10^{-4} s^{-1}	–	–
Glu_ext → Glu_astrocyte	–	–	6.367 x 10^{-12} M/s	6.1 x 10^{-5} M

Result and Discussion

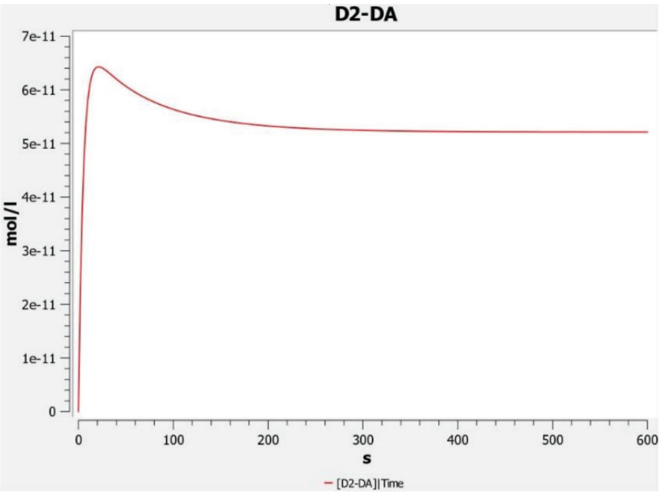


Figure 2: Graph of Trial 1 concentration of DA bound to D2 receptor over 600 seconds. This trial depicts the concentration of dopamine binding affinity with the simulated "normal" astrocyte.

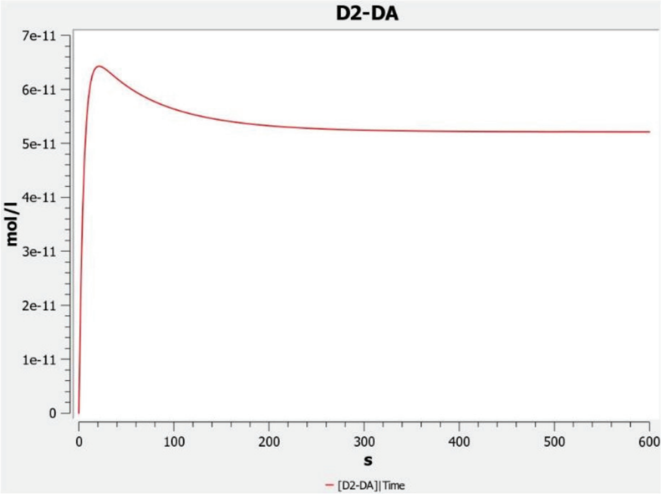


Figure 3: Graph of Trial 2 concentration of DA bound to D2 receptor over 600 seconds. This trial depicts the concentration of dopamine binding affinity to the dopamine receptor with the simulated "defective" astrocyte.

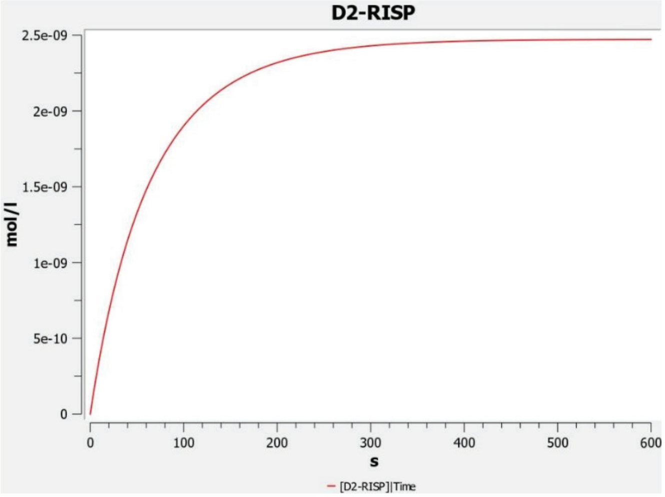


Figure 4: Trial 1 concentration of Risperidone bound to D2 receptor over 600 seconds. This trial depicts the concentration of risperidone binding affinity to the dopamine receptor with the simulated "normal" astrocyte.

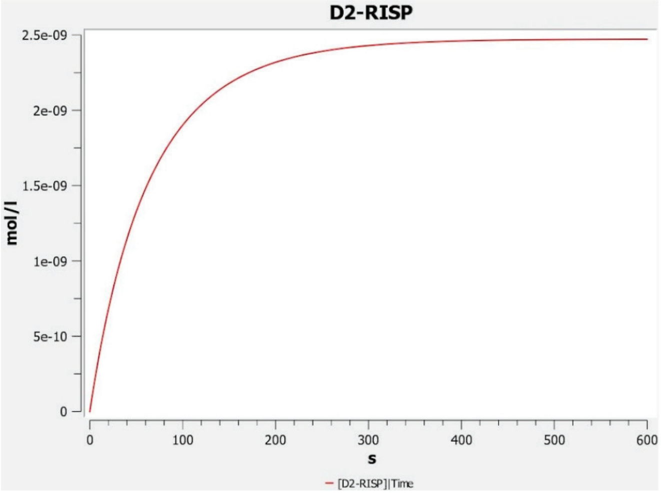


Figure 5: Trial 2 concentration of Risperidone bound to D2 receptor over 600 seconds. This trial depicts the concentration of risperidone binding affinity to the dopamine receptor with the simulated "defective" astrocyte.

DOI: 10.36838/v8i2.88

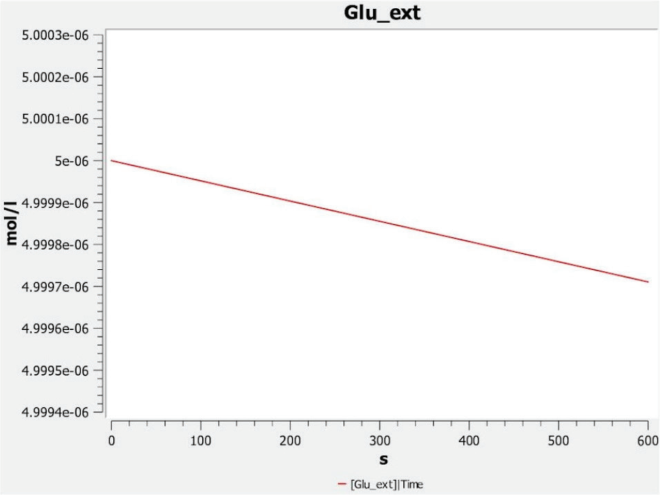


Figure 6: Trial 1 concentration of extracellular glutamate. The decrease in concentration over time of extracellular glutamate implies that the simulation of the "normal" astrocyte is accurate.

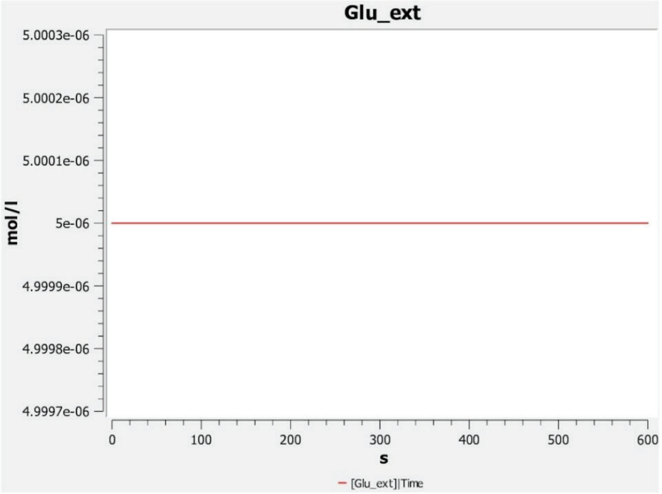


Figure 7: Trial 2 concentration of extracellular glutamate. The constant concentration of glutamate over time indicates that there is no astrocytic glutamate uptake, as intended with the simulated "defective" astrocyte.

Analysis:

The experiment aimed to evaluate the efficacy of risperidone treatment under two conditions: one with a normal Vmax for glutamate and another with Vmax set to zero. Data was collected over a period of 600 seconds for both trials, and key variables such as risperidone binding (RISP), dopamine receptor activation (D2-DA), and glutamate concentrations were analyzed. Graphs depicting these variables over time were generated to compare the trends observed in each trial. A visual inspection of the graphs for D2-DA suggests that there was no significant difference between the two trials, indicating that modifying the rate of glutamate uptake did not significantly impact dopamine receptor activation (Figures 2-3). Similarly, the trends show that the RISP values appear consistent between the two trials (Figures 4-7). To confirm whether there was a statistically significant difference in risperidone efficacy, a statistical analysis was conducted. A Welch's t-test was conducted comparing risperidone-D2 receptor binding levels (RISP) between two simulation conditions:

Condition A: Normal astrocytic glutamate uptake (V_{max} = 6.367 x 10^{-12} M/s)

Condition B: Impaired astrocytic glutamate uptake (V_{max} = 0 M/s)

Hypotheses:

• Null Hypothesis (H_0): There is no significant difference in risperidone-D2 receptor binding (RISP) between the normal and impaired astrocyte models.

• Alternative Hypothesis (H_1): There is a significant difference in risperidone-D2 receptor binding (RISP) between the normal and impaired astrocyte models.

The sample size was equal to the number of time-point measurements for each condition. The α level is 0.05 (5% significance level). It was found that the p-value was 1.0 and exceeded the alpha level. As such, the null hypothesis is not rejected. This means that there is no statistically significant difference in risperidone binding with the D2 receptor in either condition, despite the changes in glutamate uptake.

Discussion:

The results indicate that altering the Vmax of glutamate did not significantly affect risperidone binding or dopamine receptor activation over a 600-second timescale. This suggests that glutamate uptake in astrocytes does not play a major role in risperidone's mechanism of action, or it may be working in conjunction with other mechanisms. Risperidone primarily functions as a dopamine D2 and serotonin 5-HT2A receptor antagonist, and these findings support the idea that its efficacy is likely not associated with glutamate transport.

One explanation for the lack of difference between trials is that the timescale may not be sufficient to observe downstream effects of glutamate modulation. While receptor binding occurs rapidly, changes in neurotransmitter dynamics may take longer to be apparent. Additionally, the experimental model may not fully capture the complexity of glutamatergic signaling in a biological system. If glutamate uptake influences risperidone efficacy, it may be through mechanisms that were not reflected in this study.

■ Conclusion

The findings show that risperidone's efficacy remains unaffected by alterations in glutamate uptake. The statistical analysis revealed no significant variation in dopamine receptor activation or risperidone binding between trials (t = 0.0, p = 1.0), suggesting that its primary mechanism of action is not related to glutamate transport. These results contribute to a better understanding of risperidone's pharmacological effects and provide evidence that glutamate uptake does not significantly impact its therapeutic function.

Future studies should investigate whether sustained alterations in glutamate uptake over longer periods influence risperidone's effects. Additionally, utilizing in vivo experiments could better depict the nuances in brain chemistry that contribute to changes in glutamate homeostasis and play a role in risperidone's long-term efficacy.

Another area for further investigation is the interaction between glutamate uptake and other neurotransmitter systems. Since schizophrenia involves disruptions in dopamine, serotonin, and GABA pathways, studying how risperidone's effects change under different neurochemical conditions may help in developing more patient-specific treatment. Computational models could be expanded with additional receptor interactions, and additional experimentation may determine whether alternative pathways influence risperidone's clinical outcomes.

References

1. Hodkinson, A., Heneghan, C., Mahtani, K. R., Kontopantelis, E., & Panagioti, M. (2021). Benefits and harms of risperidone and paliperidone for treatment of patients with schizophrenia or bipolar disorder: A meta-analysis involving individual participant data and clinical study reports. BMC Medicine, 19(1). https://doi.org/10.1186/s12916-021-02062-w
2. Seeman, P., & Kapur, S. (2006). Schizophrenia: More dopamine, more D2 receptors. Proceedings of the National Academy of Sciences, 103(31), 10892–10897. https://doi.org/10.1073/pnas.0604311103
3. May, M., Beauchemin, M., Vary, C., Barlow, D., & Houseknecht, K. L. (2019). The antipsychotic medication, risperidone, causes global immunosuppression in healthy mice. PloS One, 14(6), e0218937. https://doi.org/10.1371/journal.pone.0218937
4. Bobermin, L. D., da Silva, A., Souza, D. O., & Quincozes-Santos, A. (2018). Differential effects of typical and atypical antipsychotics on astroglial cells in vitro. International Journal of Developmental Neuroscience, 69(1), 1–9. https://doi.org/10.1016/j.ijdevneu.2018.06.001
5. Danbolt, N. C. (2001). Glutamate uptake. Progress in Neurobiology, 65(1), 1–105. https://doi.org/10.1016/S0301-0082(00)00067-8
6. Parpura, V., Alexei, V. (2012). Astrocytes revisited: concise historic outlook on glutamate homeostasis and signaling. https://hrcak.srce.hr/file/139818
7. Butt, A. M., & Rivera, A. D. (2021). Astrocytes in bipolar disorder. Advances in Neurobiology, 26, 95–113. https://doi.org/10.1007/978-3-030-77375-5_5
8. Dai, N., Jones, B., & Husain, M. I. (2022). Astrocytes in the neuropathology of bipolar disorder: Review of current evidence. Brain Sciences, 12(11), 1513. https://doi.org/10.3390/brainsci12111513
9. Alvarez-Herrera, S., Rosel Vales, M., Pérez-Sánchez, G., Becerril-Villanueva, E., Flores-Medina, Y., Maldonado-García, J. L., Saracco-Alvarez, R., Escamilla, R., & Pavón, L. (2024). Risperidone decreases expression of serotonin receptor-2A (5-HT2A) and serotonin transporter (SERT) but not dopamine receptors and dopamine transporter (DAT) in PBMCs from patients with schizophrenia. Pharmaceuticals, 17(2), 167. https://doi.org/10.3390/ph17020167
10. Butt, A. M., & Rivera, A. D. (2021). Astrocytes in bipolar disorder. Advances in Neurobiology, 26, 95–113. https://doi.org/10.1007/978-3-030-77375-5_5

Author

Aishwarya Ashok is a high school researcher and aspiring physician recognized for award-winning neuroscience work. She earned top honors with her research at Pennsylvania Junior Academy of Science (PJAS) and Bucks County Science Research Competition (BCSRC), showcasing excellence in scientific inquiry. With a deep curiosity in neuropsychiatric conditions, Aishwarya plans to expand the depth of the current scientific understanding of behavior through innovative research.

IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Modern inside into Cutaneous Melanomagenesis

Aisha Torgautova*

1) Crimson Global Academy, 18 Stanley Street, Auckland Central, Auckland 1010, New Zealand
2) International School of Almaty, 40B Satpaev Street, Almaty, 050000, Republic of Kazakhstan; aishatorg@gmail.com*
Mentor: Dr. Hamidreza Shaye[1], Tailor Hailstock[2]

ABSTRACT: Cutaneous Melanoma is a malignant, dangerous tumor that develops from melanocytes - the only cells that synthesize melanin and accumulate it in the skin, hair follicles, and retinal pigment epithelium. Melanin provides pigmentation to the skin, eyes, and hair. This substance also absorbs harmful UV rays (ultraviolet rays) and protects cell DNA from sun damage and possible further DNA sequencing mutations it may cause. Due to a combination of genetic and environmental factors, some melanocytes may undergo malfunction in their genetic apparatus, which leads to their uncontrolled division and proliferation, eventually turning into a malignant tumor. In cutaneous melanoma, mutations in the genes *BRAF* and *NRAS* most commonly predominate among other gene mutations found during melanomagenesis, accounting for 60 percent and 20 percent, respectively. This review aims to study melanomagenesis from the perspective of a wide range of internal and external factors, their impact on gene alteration, with a detailed examination of the mutated *BRAF*, *NRAS* genes, aberrant signaling pathways, and their roles in malignant tumor formation. Moreover, this review will discuss potential melanoma therapy by directly targeting mutated genes and propose some suggestions for further drug development.

KEYWORDS: Genetics and Molecular Biology of Disease, Cutaneous Melanomagenesis, *BRAF*/*NRAS* Genes, Signaling Pathways, Targeted Therapy.

Introduction

Cutaneous Melanoma is one of the most malignant forms of skin cancer that targets both men and women, but varies by age and different risk factors.[1] Cases of melanoma have significantly increased in recent decades and continue to represent one of the most life-threatening health conditions. The National Cancer Institute estimates the mortality rate will increase by 65% and the total number of skin cancer cases to surpass 2.3 million worldwide in 2040.[2]

Cutaneous Melanoma spreads from the epidermis, and when evolving, it penetrates the dermis and subcutaneous tissue. Thus, it grows through all layers of skin - epidermis, dermis, and hypodermis. Anatomically, these tissues are very well supplied with blood and lymphatic vessels, as well as nerves. This is why Cutaneous Melanoma is mostly characterized by aggressive, fulminant development and further rapid metastasis.

Melanomas are mainly caused by gene alterations, and most of them have potentially active mutations in genes such as *BRAF* and *NRAS*. *BRAF* gene mutations account for more than 60% of all cases of Cutaneous Melanoma,[3,4] when NRAS-mutated melanomas occur in up to 15-20 % of all recorded cases.[5] NRAS-mutated melanomas are more aggressive and associated with a poorer survival prognosis compared to melanomas without *NRAS* mutation.[5] The trigger for the occurrence of Cutaneous Melanoma is sporadic mutations in genes, in addition to genetic predisposition and known risk factors associated with it.

The combination of internal and external factors triggers the evolution of morphologically and phenotypically diverse clusters of mutated melanocytes that develop Cutaneous Melanoma. The two known signaling pathways involved in malignant formation play a crucial role in uncontrolled tumor cell division, proliferation, survival, and metastasis. The therapeutic approaches used today in melanoma management are focused on targeting mutated genes that operate dysregulated signaling pathways to stop tumor progression.

1. Melanoma:

1.1. Cutaneous Melanoma:

Cutaneous Melanoma is one of the deadliest forms of all types of skin cancer and accounts for 80% of the mortality rate among all of them.[2] There are four major morphological subtypes of Cutaneous Melanoma: Superficial spreading melanoma (SSM), Lentigo melanoma (LM), Nodular Melanoma (NM), and Acral Lentiginous Melanoma (ALM). [2,6] They differ by clinical appearance and histological features, along with diagnostic biomarkers, propensity for rapid metastasis, survival rates, and treatment approaches. The data shows the number of reported cases corresponding to SSM (70%), LM (4 –15%), NM (5%), and ALM(2–5%) of the reported cases. [2,7]

Cutaneous Melanoma results from mutations in melanocytes - pigment-producing cells that are present in the stratum basale of epidermis. It is a basic single row layer of skin cells called keratinocytes, which physiologically undergo continuous cell division and therefore promote skin cell renewal. The major function of melanocytes is the production of melanin, which, when consumed by keratinocytes, forms a shield above the cell's nucleus to protect its genetic material. During embryogenesis, these two types of cells present in the skin derive from two different embryonic origins. Keratinocytes origi-

DOI: 10.36838/v8i2.93

nate from the surface ectoderm, a superficial layer that builds epithelial tissues, whereas melanocytes are formed from a multipotent stem cell of the neural crest. This difference in embryonic origin gives melanocytes "special abilities" or inbuilt potential to express many signaling molecules and factors that promote rapid invasion, migration, and propensity to rapidly metastasize to other organs when not promptly addressed.[8] Gene mutations caused by alignment of many factors result in uncontrolled cellular proliferation, tumor formation, and fulminant metastasis after malignant transformation.

Another crucial feature that characterizes Cutaneous Melanoma is its heterogeneity due to the tumor transformation of melanocytes to form genetically divergent subpopulations of cells with different morphological and phenotypic profiles composing the tumor. These subpopulations are present in the form of a small fraction of cancer stem-like cells (CSCs), responsible for the promotion of melanoma progression, drug resistance, and recurrence, and many non-cancer stem-like cells (non-CSCs) that play supportive and regulatory roles in melanogenesis.[9]

1.2. Epidemiology of Cutaneous Melanoma:

According to the American Cancer Society, Melanoma has become the third most common type of skin cancer and the fifth among all types of cancer in the US. The same source predicts that 100,640 new cases of melanoma will be diagnosed in 2024 (58,8% of men and 41,2% of women), with the expected 8,290 deaths (65,5% and 34,5%, respectively).[10,11] Over the two decades, the observed rate of Melanoma has increased from 18,1 in 2000 to 23,8 (per 100,000 population) in 2021 (Figure 1), based on data published by the National Cancer Institute (USA). At the same time, it demonstrates the lowering of the death rate from 2,7 to 2,0, respectively.[10]

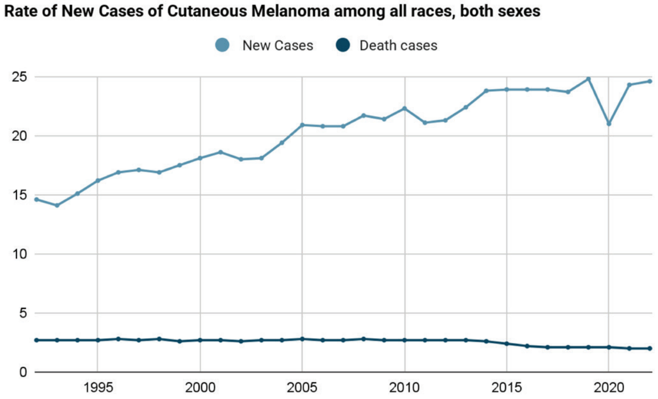
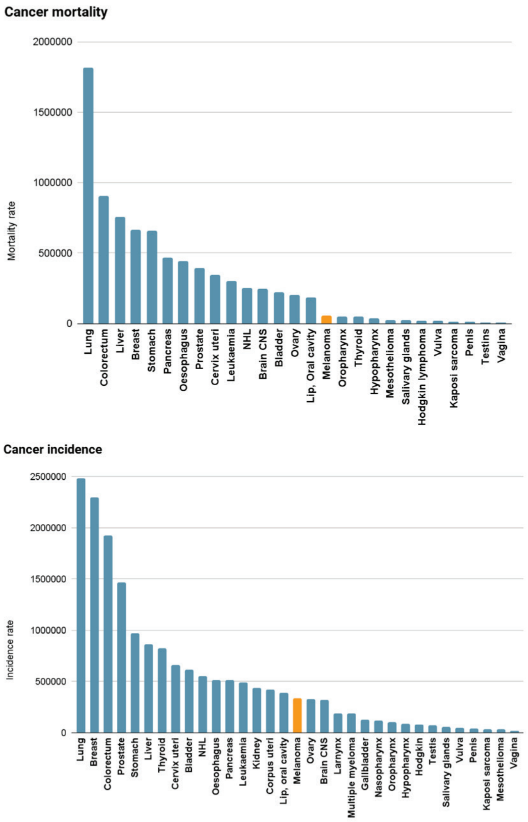


Figure 1: Rate of New Cases of Cutaneous Melanoma among all races, both sexes. The steady growth of new cases of Cutaneous Melanoma has been noted since the early 90s, while the death rate has stayed mainly unchanged. Data was acquired from the National Cancer Institute.[10]

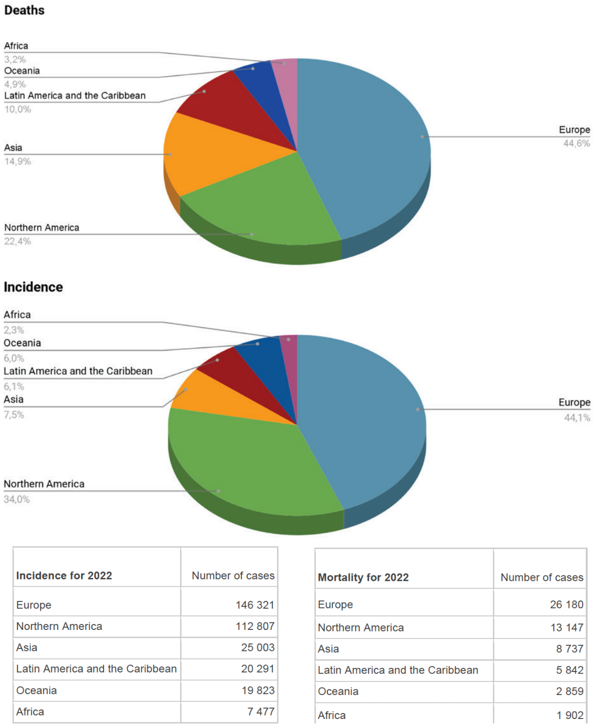
From a global perspective, Cutaneous Melanoma has become a life-threatening condition for an increasing number of people. In 2022, it ranked in the top 20 of the most common types of cancer and caused 58667 deaths.



Cutaneous Melanoma	Incidence	Mortality
Rank	17	22
Statistics	331 722	58 667

Figure 2: Cancer incidence and Mortality statistics. Among other types of cancers in 2022, the incidence and mortality rates of Cutaneous Melanoma have the 17th and 22nd ranking, respectively. Data was acquired from the World Health Organization.[12]

Worldwide statistics for 2022 show the incidence rate of Melanoma varies depending on geographical location and ethnicity. When analyzing the diagrams below (Figure 3), it becomes evident that Melanoma is prevalent among populations of European and North American countries, where the Caucasian population dominates among other ethnic groups. Altogether, these two regions come up to 78,1% (259,128) of all cases of cutaneous melanoma determined in 2022. The mortality data show lower indices in the North American region, in contrast to the other regions. This fact may indicate good diagnostic procedures that allow early melanoma detection, population awareness, and/or accessibility to modern, effective treatment schemes.



Incidence for 2022	Number of cases
Europe	146 321
Northern America	112 807
Asia	25 003
Latin America and the Caribbean	20 291
Oceania	19 823
Africa	7 477

Mortality for 2022	Number of cases
Europe	26 180
Northern America	13 147
Asia	8 737
Latin America and the Caribbean	5 842
Oceania	2 859
Africa	1 902

Figure 3: Incidence and Mortality Statistics per Region. On the worldwide scale, Europe and North America demonstrate the highest incidence and mortality rates of Cutaneous Melanoma with 44.1% and 34.0% of new cases and 44,6% and 22,4% of deaths in 2022, respectively. Data was acquired from the World Health Organization.[12]

1.3. Risk factors:

For the manifestation of Melanoma, there are a few important risk factors that may play a crucial mutagenic role in its development. It is also important to emphasize that having risk factors in addition to genetic predisposition highly increases the possibility of developing any pathological condition, including melanomagenesis.

- UV radiation

The major risk factor associated with Cutaneous Melanoma is exposure to ultraviolet (UV) rays, such as solar and artificial UV radiation. Here, it is important to note that sunburn history (especially in childhood) has a much higher Relative Risk (RR) of 2,03 in comparison with Intermittent sun exposure (RR=1.61) and Sunbathing ('ever' intentional sun exposure; RR=1.44). It can be explained by the fact that children's skin is thin, has less protective melanin, and when it faces aggressive sunlight, sun-damaged cell DNA structures form permanent mutations due to immature DNA repair systems. Accumulating additional gene mutations over the lifetime, in addition to other external factors, significantly increases the risk of cutaneous melanoma development. [13-16]

- Phenotype

Caucasian ethnicity brings a higher risk of melanoma development, especially when this factor coincides with other ones. Therefore, people with lighter skin color, in addition to having red/blond hair, blue/green eyes, and skin that freckles and burns easily, are at increased risk. It is supported by the data from the WHO incidence rate above, which demonstrates 78% of all cases detected in 2022 in North American and European regions (Figure 2), with the ethnical prevalence of the caucasian population. It may be caused not only by the amount of melanin in the skin, but also by the type of melanin produced by melanocytes. It is known that there is a type difference in melanin presented in a darker or lighter skin - Eumelanin (dark pigment) and Pheomelanin (red/yellow pigment). It is also proven that Eumelanin has a higher UV protective potential in comparison to Pheomelanin.[16,19]

- Lifestyle habits

Consumption of liquors and spirits showed to be significantly correlated with melanoma, with the highest intake (>3.08 g/day) associated with a 47% increased melanoma risk compared with the lowest intake (0–0.13 g/day).[13] Drinking alcohol can make the skin more sensitive to sunlight, decrease skin immunity, increase the toxic burden of alcoholic metabolites and oxidative products, causing gene mutations that lead to elevated vulnerability to skin cancer.[17,19]

- Immunosuppression

A decrease in immunity in general affects the state of the body and its ability to withstand external challenges, since immune surveillance of external and internal environmental factors is weakened. In this regard, the risk of cutaneous melanoma development and its resistance to immunotherapy also increases along with other pathological conditions.[18,19] Thus, immunodeficiency present in patients with HIV in Caucasians, Transplant recipients (renal), Non-Hodgkin's lymphoma, and Chronic lymphocytic leukemia correspond to the following rates of Cutaneous Melanoma: IR > 10-fold increased, RR: 3.6, RR: 2.4, and RR: 3.1, respectively.[13]

- Age

Based on the data presented by the National Cancer Institute (USA), the average age of the most frequently diagnosed skin melanoma among people is 65-74 (Figure 4), with the mean age of 66 years from 2017-2021.[10]

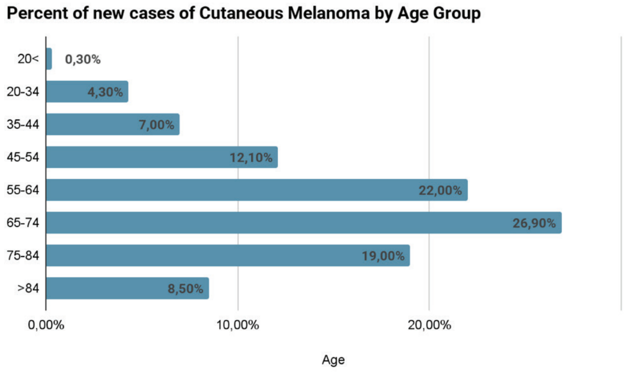


Figure 4: Percent of new cases of Cutaneous Melanoma. It is observed that the majority of new cases of Cutaneous Melanoma are mainly found in the age group of 65-74 years. Data was acquired from the National Cancer Institute.[10]

In addition, another source demonstrates a wider time period of 1992-2011 and proves the same melanoma manifestation rate of 55 years old 53%.[20] The undeniable correlation between

DOI: 10.36838/v8i2.93

older age and the occurrence of cutaneous melanoma development may account for the conclusion that melanomagenesis is a process of accumulating gene damage caused by many internal and external factors over a lifetime before turning normal melanocytes into malignant transformation.

- Sex/Gender

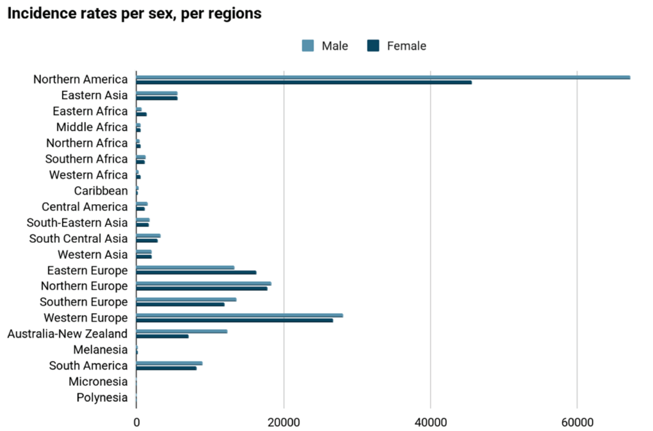


Figure 5: Incidence rates of Cutaneous Melanoma per sex, per region. It is demonstrated that on a worldwide scale, males in comparison with females are more prone to getting melanoma. Data was acquired from the World Health Organization. [12]

When studying global statistics from the perspective of sex/gender ratio, the data shows that males are more susceptible to Cutaneous Melanoma development. According to Global Cancer Observatory melanoma statistics, North America, along with Australia / New Zealand regions, demonstrates the biggest spread of incidence level between the two genders, where males are more often diagnosed with melanoma than females (Figure 5).[12] The possible conclusions that can be drawn from these statistics, excluding factors related to both sexes, are factors typical for men, such as risky sunbathing/sun exposure behavior, a higher level of alcohol intake, less awareness or melanoma alertness, and a lower conscious approach to health as a whole.

- Genetic Predisposition

Another crucial factor is genetic predisposition, which can be explained by the gene mutation shared between family members, as well as the common lifestyle habits and the same family phenotype. The risk of melanoma is higher if one or more of the first-degree relatives (parents, brothers, sisters, or children) have had melanoma or familial atypical multiple mole and melanoma (FAMMM) syndrome. Around 10% of people with melanoma have a family history of the disease.

Inherited *BRAF* and *NRAS* somatic gene mutations are characterized by incomplete penetrance, predisposing to melanoma formation, meaning that more elements need to accumulate to bring the disease to manifestation. Mutations in the *CDKN2A* gene are rare in sporadic cases but have been implicated in up to 30% of hereditary melanomas.[21] Since melanomagenesis is a multifactorial process, it requires specific genetic, epigenetic, and additional risk factors to coincide and be accumulated in one organism. The increased risk of melanomagenesis may include a shared family lifestyle of frequent sun exposure, a family tendency to have lighter skin tone, certain gene changes (mutations) that run in a family, or a combination of these factors.[12]

- Moles (Nevi)

Moles, or nevi, are benign growths of melanocytes considered to have both direct precursors and markers of increased risk for melanoma. People with >100 moles are at a seven-fold increased risk of developing melanoma in comparison to those with <15.[22] Guidelines suggest these moles should be constantly surveyed based on the ABCDE criteria (asymmetry, border irregularity, color variation, diameter >6 mm, and evolution), and if suspected, surgically removed with margins of at least 2 mm.[23]

2. Gene mutations:

2.1. Melanomagenesis:

Melanoma is a tumor composed of cells with different morphological and phenotypic profiles that form the basis for the phenomenon known as tumor heterogeneity. All these morphologically and phenotypically diverse cells are derived from normal melanocytes that have been pathologically transformed during melanomagenesis.

The genetically divergent subpopulations of tumor cells are represented by a small fraction of **CSCs** and many **non-CSCs**. These two malignant cell types differ by their stemness abilities, proliferative potential, differentiation, plasticity, metastatic activity, as well as treatment response. Other cells that are usually contributed to the normal skin morphology, such as keratinocytes, fibroblasts, endothelial and different immune cells, also play a vital role in tumor formation by releasing signaling molecules, growth factors, and cytokines that enable tumor formation and metastatic activity.

Melanoma tumor CSCs, also called Melanoma stem-like cells (MSC), are characterized by stemness properties-dependent protein markers - unique surface proteins associated with aberrant signaling pathways employed during tumor progression, drug resistance, and relapse. As a result of their origin, these cells have evolved genetically to evade drug toxicity and to promote tumor progression and metastasis. This feature may also explain why most available therapeutic approaches targeting MSCs tend to fail, and melanoma continues to proliferate and expand by spreading metastasis to nearby lymph nodes and other parts of the body.[24,25]

Many functional genes, such as ***BRAF***, *CDKN2A*, ***NRAS***, TP53, and NF1, are significantly altered by different mutations and associated with melanoma.

Among them, the most common are *BRAF* and *NRAS*, with ***BRAF***$_{V600E}$ alteration found in 50% of all melanoma cases.[9]

Since Melanoma is a tumor with heterogeneity, it demonstrates high levels of biological complexity during Melanogenesis. Consequently, melanoma cells undergo genetic, epigenetic, and/or phenotypic modification to survive in the human body.

Epigenetic alterations may play a crucial role in melanomagenesis, as these modifications in the cell genome without changing its DNA sequence may regulate gene activity through DNA methylation, histone modification, non-coding microRNA activity, or chromatin remodeling. As a result of this, certain DNA sequences, encoding certain proteins, can be turned on/off, altering their functional task.

Therefore, in order to understand Melanoma development, possible pathway activations, and its response to the applied drug treatment procedures, gene mutations should be examined carefully.

2.2. BRAF mutation:

BRAF is one of the most important genes related to melanoma formation, as more than 60% of all cutaneous melanoma cases have been proven to have mutations in this specific gene. *BRAF* is a member of the *RAF* kinase family, which plays a significant role in the regulation of essential physiological cell functions. The *BRAF* gene is located on chromosome 7 (7q34) and encodes the BRAF protein, a 94 kDa intracellular enzyme of 766 amino acids. It is involved in the Mitogen-Activated Protein Kinase/Extracellular Signal-Regulated Kinase (MAPK/ERK) signaling pathway. The MAPK/ERK pathway consists of a chain of intracellular proteins that regulates normal cell growth, differentiation, proliferation, and apoptosis.[26,27]

In other words, the final goals of this signaling pathway in physiological conditions are the control of cell cycle progression and the regulation of their life cycle through apoptosis.[28]

Single-point mutations can turn *BRAF* into an oncogene that is found predominantly in cutaneous melanoma.[29] Among of all cases of Cutaneous Melanoma caused by *BRAF* gene mutations, more than 90 % are taking place at codon 600.[3,4] At this point, single nucleotide mutation (*BRAF*V600E: nucleotide 1799 T > A; codon GTG > GAG changes aminoacid encoding from **valine (V) to glutamic acid (E)**, and results in a **480-fold** increase in BRAF protein kinase activity compared with its native form.[26,30]

Another most common mutation at codon 600 is *BRAFV600K*, substituting **valine (V)** for **lysine (K)**, 10-20 % (GTG > AAG).[31] Other rare two-nucleotide variation of the predominant mutation are *BRAFV600R* (GTG > AGG) (<5%), *BRAF V600 'E2' (GTG > GAA)* (<1%), and *BRAFV600D (GTG > GAT)* (<5%), *V600M* (<1%) and *V600G* (<1%).[26, 32-36] It is important to emphasize that both single and two-nucleotide mutations significantly affect the kinase functionality, causing cell mutagenic activity to drastically increase. The prevalence of *BRAFV600K* has been reported as being higher in some populations. Thus, *V600K* activating mutations were more common than previously reported and occurred at a rate of 20% in the Australian population that has chronic UV exposure. [35,37]

Exposure to ultraviolet light is a major causative factor in melanoma, although the relationship between risk and exposure is complex. For example, in light-skinned people, the group that is predominantly affected by melanoma, tumors are most common on areas that are intermittently exposed to the sun, such as the trunk, arms, and legs, rather than on areas that are chronically exposed to the sun, such as the face. In melanoma, tumors arising in non-sun-exposed areas and intermittently exposed to ultraviolet radiation (UVR) skin demonstrate mainly *BRAF* and *NRAS* mutations. Melanomas on chronically sun-damaged skin exhibit multiple gene mutations, where the frequency of the *BRAF* mutation declines and becomes rare.[38-40] *BRAF* mutations in cutaneous melanoma are most common on the trunk (affecting the head and neck less frequently), on skin without marked solar elastosis, and at a younger age, thus suggesting a pathophysiology role for intermittent UV exposure in early life rather than chronic sun damage.[26, 41] BRAF-mutated melanomas arise early in life at low cumulative UV doses, whereas melanomas without *BRAF* mutations require accumulation of high UV doses over time. *BRAF* mutations in cutaneous melanoma are independently associated with age, anatomic site of the primary tumor, and the degree of solar elastosis at the primary tumor site.[41]

2.3. NRAS mutation:

NRAS mutations are found in 15%–20% of melanomas. The *NRAS* gene mutation was the first oncogene identified in melanoma in 1986.

The *NRAS* gene is located on chromosome 1 (1p13.2) and encodes the protein *NRAS* that acts as a GTPase and plays the role of a molecular switch between active and inactive states. Commonly, *NRAS* mutations are found at codons 12, 61, or, less frequently, 13, and represent single-nucleotide mutations.[42,50]

Whereas a mutant *NRAS(Q61)* gene disrupts the GTPase activity of RAS, locking it in its active conformation, leading to continuous activation of the MAPK/ERK pathway. *NRAS(G12)* and *NRAS(G13)* mutations contribute to the structural changes in the protein, thus decreasing its sensitivity to GTPase-accelerating proteins and also resulting in sustained activation of the MAPK/ERK pathway, although to a lesser degree. PI3K/AKT (phosphoinositide 3-kinase) is another pathway that can be altered by the *NRAS* gene mutation.[43, 44]

Typical patients harboring the *NRAS* mutation tend to be older (over 55) and have a history of chronic ultraviolet (UV) exposure.[45-47] The lesions are usually located at the extremities and have greater levels of mitosis than BRAF-mutant melanomas. Moreover, *NRAS* mutations are associated with lower rates of ulceration and thicker primary tumors. Histologically, mutant *NRAS* tumors are more aggressive than other subtypes, have elevated mitotic activity, and have higher rates of lymph node metastasis.[48-50]

2.4. Signaling pathways:

Signaling pathways play an important role in the regulation of many biological processes. Studying and a deep understanding of their perplexing mechanisms, alterations that eventually replace normal physiological conditions due to the malfunction of mutated genes, offers a way to develop efficient targeted therapeutic approaches to the treatment of cutaneous melanoma.

MAPK/ERK pathway:

A special role in the process of melanogenesis is played by the mitogen-activated protein kinase (MAPK) pathway, which regulates cell survival, proliferation, differentiation, apoptosis, and cell response to different stress factors.

The MAPK signaling pathway is a complex network of signaling cascades, including several branches with different functional characteristics and biological consequences.

One of the powerful activators of MAPK signaling pathways is epidermal growth factor (EGF).[57] It activates the MAPK cascade, starting with phosphorylation and activation of kinases at the top of the cascade, such as RAF, MEK, and ERK. This process leads to signal transduction from cell surface receptors to target sites inside the cell, initiating various cellular responses such as growth, proliferation, differentiation, and survival.

Phosphorylation and dephosphorylation determine their functional activity inside the cell, playing a key role in regulating the activity of MAPK kinases.

Dephosphorylation of MAPK kinases ensures the shutdown of the signaling cascade and the return of the cell to the baseline level of activity. This process can be carried out by various enzymes (phosphatases), which remove phosphate groups, thereby reducing the kinase activity of MAPK. Like phosphorylation, dephosphorylation is a key mechanism for regulating cellular signaling pathways. Dephosphorylation allows the cell to precisely control its responses to external signals and maintain homeostasis within the cellular environment. This balance between phosphorylation and dephosphorylation significantly affects cellular functions and the general condition of the organism. Understanding these mechanisms is important for the development of new treatments and diagnostics for many diseases associated with dysfunction of cell signaling.[58]

Mutations in the *BRAS* and *NRAS* genes turn on the ligand-independent activation of MEK or Mitogen-Activated Protein Kinase Kinase, also known as MAP2K, bypassing prior binding of epidermal growth factor (EGF) to its receptor (EGFR) on the cell membrane that normally leads to the activation of MEK. The whole MAPK/ERK pathway represents a cascade of biochemical reactions during which several key proteins are being activated in a sequence: RAS-RAF-MEK-ERK. The last one translocates signals to the cell nucleus, promoting cell proliferation, differentiation, and survival. Dysregulated of gene mutation MAP/ERK pathway results in uncontrolled cell growth and the formation of tumors, as well as the inhibition of apoptosis.[51]

PI3K/AKT/mTOR pathway:

Another crucial pathway employed in melanomagenesis is PI3K/AKT/mTOR. It plays an important role in cell growth, survival, and metabolism. Like the MAPK signaling pathway, the PI3K/Akt/mTOR pathway is activated in response to extracellular signals, such as growth factors and cytokines, through the activation of tyrosine kinase receptors (RTKs) and other cell surface receptors. Activated PI3K/AKT/mTOR pathway through a series of biochemical reactions leads to the activation of mTORC1 that promotes protein synthesis, cell growth, and survival.

However, when dysregulated, this pathway is often observed in melanoma due to genetic alterations, such as mutations and amplifications in pathway components.[51,52]

This pathway, along with cell growth and survival, confers resistance to applied therapies.

The two pathways - **MAPK/ERK** and **PI3K/AKT/mTOR** (Figure 6) often interact and may be simultaneously activated. It can contribute to more aggressive tumor cell proliferation, development, and treatment resistance.[51]

The scheme presented in Figure 6 demonstrates the key components of both pathways that are initiated at the same site and through the alternative biochemical reactions lead to the same biological response.

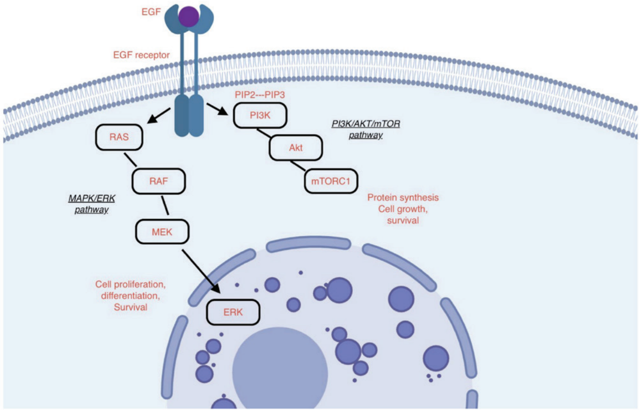


Figure 6: Signaling pathways involved in *BRAF* and *NRAS* gene mutations. The main role in melanomagenesis is played by two pathways- MAPK/ERK and PI3K/ERK, which often can interact and may be simultaneously activated, leading to more aggressive tumor cell proliferation and treatment resistance. (made in https://www.biorender.com/)

3. Therapies:

The earlier applied therapeutic methods in addressing melanoma were limited to Surgical Resection, Chemotherapy, Radiation therapy, and Immunotherapy. Nowadays, the traditional approach to melanoma treatment is being extensively developed in the direction of Targeted Therapy. It is focused on reaching the dysregulated pathways of genes involved in cell growth, their differentiation, and functionality. Melanomagenesis and its further development are mediated by genetic and epigenetic alterations amplified by the variety of risk factors that make changes to the multiple signaling pathways, including MAPK/ERK, PI3K/AKT/mTOR, and other ones not mentioned in this paper (JNK and Jak/STAT pathways).[9, 53, 54]

However, the biggest challenge and main issue in targeted therapy presented by inhibition of mutated *BRAF* and *NRAS* genes turned out to be Drug resistance and Melanoma recurrence.

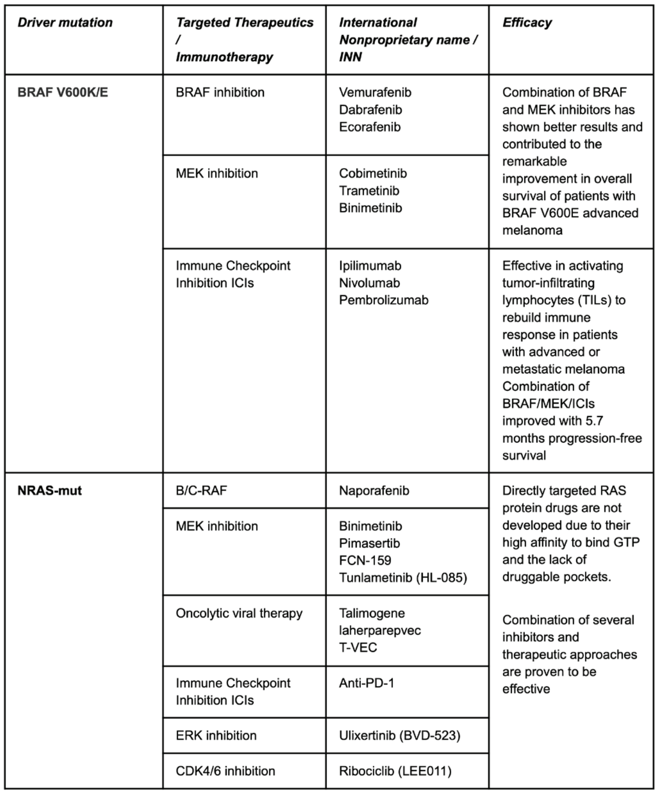
It has been stated that melanoma progression and treatment failures are attributed to tumor heterogeneity due to genetically divergent subpopulations – CSCs and non-CSCs. The stemness property of CSCs (MSCs) leads to increased drug metabolism, enhanced repair capacity of damaged DNA, reactivation of drug targets, overactivation of growth and survival signaling pathways, amplifications, and impaired activity of apoptosis/autophagy-dependent pathways.[55,56]

Both signaling pathways studied in this review paper, MAPK/ERK and PI3K/AKT/mTOR, are interconnected at multiple points, and inhibition of one of them may not only fail to stop the development of the disease but also provoke its active growth by the activation of the other signaling cascade.

Improved clinical outcomes and treatment efficiency might be reasonably developed by the intersection of MAPK/ERK and PI3K/AKT/mTOR pathways simultaneously.[9]

The table below presents the major treatment procedures applied to the cutaneous melanoma provoked by *BRAF* and *NRAS* mutations:[59-61]

Table 1: Targeted therapies are applied for BRAF and NRAS driver mutations.



Driver mutation	***Targeted Therapeutics / Immunotherapy***	***International Nonproprietary name / INN***	***Efficacy***
BRAF V600K/E	BRAF inhibition	Vemurafenib Dabrafenib Ecorafenib	Combination of BRAF and MEK inhibitors has shown better results and contributed to the remarkable improvement in overall survival of patients with BRAF V600E advanced melanoma
	MEK inhibition	Cobimetinib Trametinib Binimetinib	
	Immune Checkpoint Inhibition ICIs	Ipilimumab Nivolumab Pembrolizumab	Effective in activating tumor-infiltrating lymphocytes (TILs) to rebuild immune response in patients with advanced or metastatic melanoma Combination of BRAF/MEK/ICIs improved with 5.7 months progression-free survival
NRAS-mut	B/C-RAF	Naporafenib	Directly targeted RAS protein drugs are not developed due to their high affinity to bind GTP and the lack of druggable pockets.
	MEK inhibition	Binimetinib Pimasertib FCN-159 Tunlametinib (HL-085)	
	Oncolytic viral therapy	Talimogene laherparepvec T-VEC	Combination of several inhibitors and therapeutic approaches are proven to be effective
	Immune Checkpoint Inhibition ICIs	Anti-PD-1	
	ERK inhibition	Ulixertinib (BVD-523)	
	CDK4/6 inhibition	Ribociclib (LEE011)	

Conclusion

Melanoma is a malignant tumor composed of genetically divergent subpopulations of cells, presented by a small fraction of CSCs and many non-CSCs. Due to tumor heterogeneity and special properties of CSCs cells, Melanoma is prone to rapid development and metastasis in different organs of the human body, which affects treatment outcomes and life prognosis. The data presented in this review paper demonstrate that different risk factors such as UV radiation, Phenotype, Age/ Gender, Lifestyle behavior, Family disease history, and Health condition, along with genetic mutations, play a crucial role in the development of Cutaneous Melanoma.

Mutations in the genes *BRAF* (60%) and *NRAS* (20%) most commonly dominate among other gene mutations found during melanomagenesis and cause dysregulation in MAPK/ERK and PI3K/AKT/mTOR signaling pathways responsible for cell growth, differentiation, and functional activities. Interconnectedness between the two major signaling pathways results in the targeted treatment failure and disease recurrence when singly addressed.

Over the past few decades, treatment options for cutaneous melanoma have advanced significantly, improving survival rates in patients with *BRAF* and *NRAS* mutations. However, not all the driver mutations are effectively targeted, especially in *NRAS* mutations.

There is a big need for new treatment procedures to address all known pathways through targeted therapy. The fact that these pathways are interconnected with each other and when one is blocked, the path can continue by an alternative route, should be taken into consideration by future researchers. Due to the limitations in available scientific data on a wider spectrum of additional signaling pathways activated during melanogenesis, further research studies for potentially effective novel therapies targeting points of pathway intersections are needed in the field of Cutaneous melanoma treatment.

Acknowledgments

I would like to thank Dr. Shaye for informative lectures and valuable feedback. I'm grateful to Taylor Hailstock, TA, for her prompt responses and also to my classmates for helpful suggestions on my review paper. I attest that the ideas, graphics, and writing in this paper are entirely my own.

References

1. Melanoma|Cancer.Net. Available online: https://www.cancer.net/cancer-types/melanoma/statistics (accessed on 13 August 2024).
2. Chatzilakou, E.; Hu, Y.; Jiang, N.; Yetisen, A. K. Biosensors for Melanoma Skin Cancer Diagnostics. *Biosensors and Bioelectronics* **2024**, *250*, 116045. https://doi.org/10.1016/j.bios.2024.116045.
3. Wellbrock, C.; Hurlstone, A. BRAF as Therapeutic Target in Melanoma. *Biochemical Pharmacology* **2010**, *80* (5), 561–567. https://doi.org/10.1016/j.bcp.2010.03.019.
4. CiVelek, Z.; Kfashi, M. An Improved Deep CNN For an Early and Accurate Skin Cancer Detection and Diagnosis System. *IJERAD* **2022**, *14* (2), 721–734. https://doi.org/10.29137/umagd.1116295.
5. Muñoz-Couselo, E.; Zamora Adelantado, E.; Ortiz Vélez, C.; Soberino-García, J.; Pérez-García, J. M. NRAS-Mutant Melanoma: Current Challenges and Future Prospect. *OTT* **2017**, *Volume 10*, 3941–3947. https://doi.org/10.2147/OTT.S117121.
6. D'Orazio, J.; Jarrett, S.; Amaro-Ortiz, A.; Scott, T. UV Radiation and the Skin. *IJMS* **2013**, *14* (6), 12222–12248. https://doi.org/10.3390/ijms140612222.
7. *Primary Care Dermatology*; Pujalte, G. G. A., Heidelbaugh, J. J., Eds.; Primary care clinics in office practice; Elsevier: Philadelphia, 2015
8. Saginala, K.; Barsouk, A.; Aluru, J. S.; Rawla, P.; Barsouk, A. Epidemiology of Melanoma. *Medical Sciences* **2021**, *9* (4), 63. https://doi.org/10.3390/medsci9040063.
9. Al Hmada, Y.; Brodell, R. T.; Kharouf, N.; Flanagan, T. W.; Alamodi, A. A.; Hassan, S.-Y.; Shalaby, H.; Hassan, S.-L.; Haikel, Y.; Megahed, M.; Santourlidis, S.; Hassan, M. Mechanisms of Melanoma Progression and Treatment Resistance: Role of Cancer Stem-like Cells. *Cancers* **2024**, *16* (2), 470. https://doi.org/10.3390/cancers16020470.

10. National Cancer Institute Melanoma of the Skin-Cancer Stat Facts. Available online: https://seer.cancer.gov/statfacts/html/melan.html (accessed on 13 August 2024).
11. National Cancer Society/ Key statistics for Melanoma Skin Cancer. Available online: https://www.cancer.org/cancer/types/melanoma-skin-cancer/about/key-statistics.html (accessed on 13 August 2024).
12. WHO/International Agency for research on cancer/ Global Cancer Observatory. Available online: https://gco.iarc.who.int/media/globocan/factsheets/cancers/16-melanoma-of-skin-fact-sheet.pdf (accessed on 13 August 2024).
13. Wunderlich, K.; Suppa, M.; Gandini, S.; Lipski, J.; White, J. M.; Del Marmol, V. Risk Factors and Innovations in Risk Assessment for Melanoma, Basal Cell Carcinoma, and Squamous Cell Carcinoma. *Cancers* **2024**, *16* (5), 1016. https://doi.org/10.3390/cancers16051016.
14. Gandini, S.; Sera, F.; Cattaruzza, M. S.; Pasquini, P.; Picconi, O.; Boyle, P.; Melchi, C. F. Meta-Analysis of Risk Factors for Cutaneous Melanoma: II. Sun Exposure. *European Journal of Cancer* **2005**, *41* (1), 45–60. https://doi.org/10.1016/j.ejca.2004.10.016.
15. O'Sullivan, D. E.; Brenner, D. R.; Villeneuve, P. J.; Walter, S. D.; Demers, P. A.; Friedenreich, C. M.; King, W. D. Estimates of the Current and Future Burden of Melanoma Attributable to Ultraviolet Radiation in Canada. *Preventive Medicine* **2019**, *122*, 81–90. https://doi.org/10.1016/j.ypmed.2019.03.012.
16. Nasti, T. H.; Timares, L. MC 1R, Eumelanin and Pheomelanin: Their Role in Determining the Susceptibility to Skin Cancer. *Photochem & Photobiology* **2015**, *91* (1), 188–200. https://doi.org/10.1111/php.12335.
17. Mahamat-Saleh, Y.; Al-Rahmoun, M.; Severi, G.; Ghiasvand, R.; Veierod, M. B.; Caini, S.; Palli, D.; Botteri, E.; Sacerdote, C.; Ricceri, F.; Lukic, M.; Sánchez, M. J.; Pala, V.; Tumino, R.; Chiodini, P.; Amiano, P.; Colorado-Yohar, S.; Chirlaque, M.; Ardanaz, E.; Bonet, C.; Katzke, V.; Kaaks, R.; Schulze, M. B.; Overvad, K.; Dahm, C. C.; Antoniussen, C. S.; Tjønneland, A.; Kyrø, C.; Bueno-de-Mesquita, B.; Manjer, J.; Jansson, M.; Esberg, A.; Mori, N.; Ferrari, P.; Weiderpass, E.; Boutron-Ruault, M.; Kvaskoff, M. Baseline and Lifetime Alcohol Consumption and Risk of Skin Cancer in the European Prospective Investigation into Cancer and Nutrition Cohort (EPIC). *Intl Journal of Cancer* **2023**, *152* (3), 348–362. https://doi.org/10.1002/ijc.34253.
18. Mahmoud, F.; Shields, B.; Makhoul, I.; Avaritt, N.; Wong, H. K.; Hutchins, L. F.; Shalin, S.; Tackett, A. J. Immune Surveillance in Melanoma: From Immune Attack to Melanoma Escape and Even Counterattack. *Cancer Biology & Therapy* **2017**, *18* (7), 451–469. https://doi.org/10.1080/15384047.2017.1323596.
19. National Cancer Society/Melanoma Skin Cancer Risk Factors. Available online: https://www.cancer.org/cancer/types/melanoma-skin-cancer/causes-risks-prevention/risk-factors.htm(accessed on 14 August 2024).
20. Enninga, E. A. L.; Moser, J. C.; Weaver, A. L.; Markovic, S. N.; Brewer, J. D.; Leontovich, A. A.; Hieken, T. J.; Shuster, L.; Kottschade, L. A.; Olariu, A.; Mansfield, A. S.; Dronca, R. S. Survival of Cutaneous Melanoma Based on Sex, Age, and Stage in the United States, 1992–2011. *Cancer Medicine* **2017**, *6* (10), 2203–2212. https://doi.org/10.1002/cam4.1152.
21. Berwick, M.; Erdei, E.; Hay, J. Melanoma Epidemiology and Public Health. *Dermatologic Clinics* **2009**, *27* (2), 205–214. https://doi.org/10.1016/j.det.2008.12.002.
22. Gandini, S.; Sera, F.; Cattaruzza, M. S.; Pasquini, P.; Abeni, D.; Boyle, P.; Melchi, C. F. Meta-Analysis of Risk Factors for Cutaneous Melanoma: I. Common and Atypical Naevi. *European Journal of Cancer* **2005**, *41* (1), 28–44. https://doi.org/10.1016/j.ejca.2004.10.015.
23. Terushkin, V.; Ng, E.; Stein, J. A.; Katz, S.; Cohen, D. E.; Meehan, S.; Polsky, D. A Prospective Study Evaluating the Utility of a 2-Mm Biopsy Margin for Complete Removal of Histologically Atypical (Dysplastic) Nevi. *Journal of the American Academy of Dermatology* **2017**, *77* (6), 1096–1099. https://doi.org/10.1016/j.jaad.2017.07.016.
24. El-Khattouti, A.; Selimovic, D.; Haïkel, Y.; Megahed, M.; Gomez, C. R.; Hassan, M. Identification and Analysis of CD133+ Melanoma Stem-like Cells Conferring Resistance to Taxol: An Insight into the Mechanisms of Their Resistance and Response. *Cancer Letters* **2014**, *343* (1), 123–133. https://doi.org/10.1016/j.canlet.2013.09.024.
25. Fattore, L.; Mancini, R.; Ciliberto, G. Cancer Stem Cells and the Slow Cycling Phenotype: How to Cut the Gordian Knot Driving Resistance to Therapy in Melanoma. *Cancers* **2020**, *12* (11), 3368. https://doi.org/10.3390/cancers12113368.
26. Ottaviano, M.; Giunta, E.; Tortora, M.; Curvietto, M.; Attademo, L.; Bosso, D.; Cardalesi, C.; Rosanova, M.; De Placido, P.; Pietroluongo, E.; Riccio, V.; Mucci, B.; Parola, S.; Vitale, M.; Palmieri, G.; Daniele, B.; Simeone, E.; on behalf of SCITO YOUTH. BRAF Gene and Melanoma: Back to the Future. *IJMS* **2021**, *22* (7), 3474. https://doi.org/10.3390/ijms22073474.
27. Peyssonnaux, C.; Eychène, A. The Raf/MEK/ERK Pathway: New Concepts of Activation. *Biology of the Cell* **2001**, *93* (1–2), 53–62. https://doi.org/10.1016/S0248-4900(01)01125-X.
28. McCubrey, J. A.; Steelman, L. S.; Chappell, W. H.; Abrams, S. L.; Wong, E. W. T.; Chang, F.; Lehmann, B.; Terrian, D. M.; Milella, M.; Tafuri, A.; Stivala, F.; Libra, M.; Basecke, J.; Evangelisti, C.; Martelli, A. M.; Franklin, R. A. Roles of the Raf/MEK/ERK Pathway in Cell Growth, Malignant Transformation and Drug Resistance. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **2007**, *1773* (8), 1263–1284. https://doi.org/10.1016/j.bbamcr.2006.10.001.
29. Wellbrock, C.; Hurlstone, A. BRAF as Therapeutic Target in Melanoma. *Biochemical Pharmacology* **2010**, *80* (5), 561–567. https://doi.org/10.1016/j.bcp.2010.03.019.
30. Wan, P. T. C.; Garnett, M. J.; Roe, S. M.; Lee, S.; Niculescu-Duvaz, D.; Good, V. M.; Project, C. G.; Jones, C. M.; Marshall, C. J.; Springer, C. J.; Barford, D.; Marais, R. Mechanism of Activation of the RAF-ERK Signaling Pathway by Oncogenic Mutations of B-RAF. *Cell* **2004**, *116* (6), 855–867. https://doi.org/10.1016/S0092-8674(04)00215-6.
31. Rubinstein, J. C.; Sznol, M.; Pavlick, A. C.; Ariyan, S.; Cheng, E.; Bacchiocchi, A.; Kluger, H. M.; Narayan, D.; Halaban, R. Incidence of the V600K Mutation among Melanoma Patients with BRAF Mutations, and Potential Therapeutic Response to the Specific BRAF Inhibitor PLX4032. *J Transl Med* **2010**, *8* (1), 67. https://doi.org/10.1186/1479-5876-8-67.
32. Gopal, P.; Sarihan, E. I.; Chie, E. K.; Kuzmishin, G.; Doken, S.; Pennell, N. A.; Raymond, D. P.; Murthy, S. C.; Ahmad, U.; Raja, S.; Almeida, F.; Sethi, S.; Gildea, T. R.; Peacock, C. D.; Adams, D. J.; Abazeed, M. E. Clonal Selection Confers Distinct Evolutionary Trajectories in BRAF-Driven Cancers. *Nat Commun* **2019**, *10* (1), 5143. https://doi.org/10.1038/s41467-019-13161-x.
33. Heinzerling, L.; Kühnapfel, S.; Meckbach, D.; Baiter, M.; Kaempgen, E.; Keikavoussi, P.; Schuler, G.; Agaimy, A.; Bauer, J.; Hartmann, A.; Kiesewetter, F.; Schneider-Stock, R. Rare BRAF Mutations in Melanoma Patients: Implications for Molecular Testing in Clinical Practice. *Br J Cancer* **2013**, *108* (10), 2164–2171. https://doi.org/10.1038/bjc.2013.143.

34. Akbani, R.; Akdemir, K. C.; Aksoy, B. A.; Albert, M.; Ally, A.; Amin, S. B.; Arachchi, H.; Arora, A.; Auman, J. T.; Ayala, B.; Baboud, J.; Balasundaram, M.; Balu, S.; Barnabas, N.; Bartlett, J.; Bartlett, P.; Bastian, B. C.; Baylin, S. B.; Behera, M.; Belyaev, D.; Benz, C.; Bernard, B.; Beroukhim, R.; Bir, N.; Black, A. D.; Bodenheimer, T.; Boice, L.; Boland, G. M.; Bono, R.; Bootwalla, M. S.; Bosenberg, M.; Bowen, J.; Bowlby, R.; Bristow, C. A.; Brockway-Lunardi, L.; Brooks, D.; Brzezinski, J.; Bshara, W.; Buda, E.; Burns, W. R.; Butterfield, Y. S. N.; Button, M.; Calderone, T.; Cappellini, G. A.; Carter, C.; Carter, S. L.; Cherney, L.; Cherniack, A. D.; Chevalier, A.; Chin, L.; Cho, J.; Cho, R. J.; Choi, Y.-L.; Chu, A.; Chudamani, S.; Cibulskis, K.; Ciriello, G.; Clarke, A.; Coons, S.; Cope, L.; Crain, D.; Curley, E.; Danilova, L.; D'Atri, S.; Davidsen, T.; Davies, M. A.; Delman, K. A.; Demchok, J. A.; Deng, Q. A.; Deribe, Y. L.; Dhalla, N.; Dhir, R.; DiCara, D.; Dinikin, M.; Dubina, M.; Ebrom, J. S.; Egea, S.; Eley, G.; Engel, J.; Eschbacher, J. M.; Fedosenko, K. V.; Felau, I.; Fennell, T.; Ferguson, M. L.; Fisher, S.; Flaherty, K. T.; Frazer, S.; Frick, J.; Fulidou, V.; Gabriel, S. B.; Gao, J.; Gardner, J.; Garraway, L. A.; Gastier-Foster, J. M.; Gaudioso, C.; Gehlenborg, N.; Genovese, G.; Gerken, M.; Gershenwald, J. E.; Getz, G.; Gomez-Fernandez, C.; Gribbin, T.; Grimsby, J.; Gross, B.; Guin, R.; Gutschner, T.; Hadjipanayis, A.; Halaban, R.; Hanf, B.; Haussler, D.; Haydu, L. E.; Hayes, D. N.; Hayward, N. K.; Heiman, D. I.; Herbert, L.; Herman, J. G.; Hersey, P.; Hoadley, K. A.; Hodis, E.; Holt, R. A.; Hoon, D. SB.; Hoppough, S.; Hoyle, A. P.; Huang, F. W.; Huang, M.; Huang, S.; Hutter, C. M.; Ibbs, M.; Iype, L.; Jacobsen, A.; Jakrot, V.; Janning, A.; Jeck, W. R.; Jefferys, S. R.; Jensen, M. A.; Jones, C. D.; Jones, S. J. M.; Ju, Z.; Kakavand, H.; Kang, H.; Kefford, R. F.; Khuri, F. R.; Kim, J.; Kirkwood, J. M.; Klode, J.; Korkut, A.; Korski, K.; Krauthammer, M.; Kucherlapati, R.; Kwong, L. N.; Kycler, W.; Ladanyi, M.; Lai, P. H.; Laird, P. W.; Lander, E.; Lawrence, M. S.; Lazar, A. J.; Łaźniak, R.; Lee, D.; Lee, J. E.; Lee, J.; Lee, K.; Lee, S.; Lee, W.; Leporowska, E.; Leraas, K. M.; Li, H. I.; Lichtenberg, T. M.; Lichtenstein, L.; Lin, P.; Ling, S.; Liu, J.; Liu, O.; Liu, W.; Long, G. V.; Lu, Y.; Ma, S.; Ma, Y.; Mackiewicz, A.; Mahadeshwar, H. S.; Malke, J.; Mallery, D.; Manikhas, G. M.; Mann, G. J.; Marra, M. A.; Matejka, B.; Mayo, M.; Mehrabi, S.; Meng, S.; Meyerson, M.; Mieczkowski, P. A.; Miller, J. P.; Miller, M. L.; Mills, G. B.; Moiseenko, F.; Moore, R. A.; Morris, S.; Morrison, C.; Morton, D.; Moschos, S.; Mose, L. E.; Muller, F. L.; Mungall, A. J.; Murawa, D.; Murawa, P.; Murray, B. A.; Nezi, L.; Ng, S.; Nicholson, D.; Noble, M. S.; Osunkoya, A.; Owonikoko, T. K.; Ozenberger, B. A.; Pagani, E.; Paklina, O. V.; Pantazi, A.; Parfenov, M.; Parfitt, J.; Park, P. J.; Park, W.-Y.; Parker, J. S.; Passarelli, F.; Penny, R.; Perou, C. M.; Pihl, T. D.; Potapova, O.; Prieto, V. G.; Protopopov, A.; Quinn, M. J.; Radenbaugh, A.; Rai, K.; Ramalingam, S. S.; Raman, A. T.; Ramirez, N. C.; Ramirez, R.; Rao, U.; Rathmell, W. K.; Ren, X.; Reynolds, S. M.; Roach, J.; Robertson, A. G.; Ross, M. I.; Roszik, J.; Russo, G.; Saksena, G.; Saller, C.; Samuels, Y.; Sander, C.; Sander, C.; Sandusky, G.; Santoso, N.; Saul, M.; Saw, R. PM.; Schadendorf, D.; Schein, J. E.; Schultz, N.; Schumacher, S. E.; Schwallier, C.; Scolyer, R. A.; Seidman, J.; Sekhar, P. C.; Sekhon, H. S.; Senbabaoglu, Y.; Seth, S.; Shannon, K. F.; Sharpe, S.; Sharpless, N. E.; Shaw, K. R. M.; Shelton, C.; Shelton, T.; Shen, R.; Sheth, M.; Shi, Y.; Shiau, C. J.; Shmulevich, I.; Sica, G. L.; Simons, J. V.; Sinha, R.; Sipahimalani, P.; Sofia, H. J.; Soloway, M. G.; Song, X.; Sougnez, C.; Spillane, A. J.; Spychała, A.; Stretch, J. R.; Stuart, J.; Suchorska, W. M.; Sucker, A.; Sumer, S. O.; Sun, Y.; Synott, M.; Tabak, B.; Tabler, T. R.; Tam, A.; Tan, D.; Tang, J.; Tarnuzzer, R.; Tarvin, K.; Tatka, H.; Taylor, B. S.; Teresiak, M.; Thiessen, N.; Thompson, J. F.; Thorne, L.; Thorsson, V.; Trent, J. M.; Triche, T. J.; Tsai, K. Y.; Tsou, P.; Van Den Berg, D. J.; Van Allen, E. M.; Veluvolu, U.; Verhaak, R. G.; Voet, D.; Voronina, O.; Walter, V.; Walton, J. S.; Wan, Y.; Wang, Y.; Wang, Z.; Waring, S.; Watson, I. R.; Weinhold, N.; Weinstein, J. N.; Weisenberger, D. J.; White, P.; Wilkerson, M. D.; Wilmott, J. S.; Wise, L.; Wiznerowicz, M.; Woodman, S. E.; Wu, C.-J.; Wu, C.-C.; Wu, J.; Wu, Y.; Xi, R.; Xu, A. W.; Yang, D.; Yang, L.; Yang, L.; Zack, T. I.; Zenklusen, J. C.; Zhang, H.; Zhang, J.; Zhang, W.; Zhao, X.; Zhu, J.; Zhu, K.; Zimmer, L.; Zmuda, E.; Zou, L. Genomic Classification of Cutaneous Melanoma. *Cell* **2015**, *161* (7), 1681–1696. https://doi.org/10.1016/j.cell.2015.05.044.
35. Ascierto, P. A.; Kirkwood, J. M.; Grob, J.-J.; Simeone, E.; Grimaldi, A. M.; Maio, M.; Palmieri, G.; Testori, A.; Marincola, F. M.; Mozzillo, N. The Role of BRAF V600 Mutation in Melanoma. *J Transl Med* **2012**, *10* (1), 85. https://doi.org/10.1186/1479-5876-10-85.
36. Catalogue of Somatic Mutations in Cancer (COSMIC): [http://www.sanger.ac.uk/cosmic]
37. Long, G. V.; Menzies, A. M.; Nagrial, A. M.; Haydu, L. E.; Hamilton, A. L.; Mann, G. J.; Hughes, T. M.; Thompson, J. F.; Scolyer, R. A.; Kefford, R. F. Prognostic and Clinicopathologic Associations of Oncogenic BRAF in Metastatic Melanoma. *JCO* **2011**, *29* (10), 1239–1246. https://doi.org/10.1200/JCO.2010.32.4327.
38. Soura, E.; Stratigos, A. J. Melanoma. In *European Handbook of Dermatological Treatments*; Katsambas, A. D., Lotti, T. M., Dessinioti, C., D'Erme, A. M., Eds.; Springer International Publishing: Cham, 2023; pp 623–637. https://doi.org/10.1007/978-3-031-15130-9_58.
39. Curtin, J. A.; Fridlyand, J.; Kageshita, T.; Patel, H. N.; Busam, K. J.; Kutzner, H.; Cho, K.-H.; Aiba, S.; Bröcker, E.-B.; LeBoit, P. E.; Pinkel, D.; Bastian, B. C. Distinct Sets of Genetic Alterations in Melanoma. *N Engl J Med* **2005**, *353* (20), 2135–2147. https://doi.org/10.1056/NEJMoa050092.
40. McKay, M. M.; Morrison, D. K. Integrating Signals from RTKs to ERK/MAPK. *Oncogene* **2007**, *26* (22), 3113–3121. https://doi.org/10.1038/sj.onc.1210394.
41. Bauer, J.; Büttner, P.; Murali, R.; Okamoto, I.; Kolaitis, N. A.; Landi, M. T.; Scolyer, R. A.; Bastian, B. C. BRAF Mutations in Cutaneous Melanoma Are Independently Associated with Age, Anatomic Site of the Primary Tumor, and the Degree of Solar Elastosis at the Primary Tumor Site. *Pigment Cell Melanoma Res* **2011**, *24* (2), 345–351. https://doi.org/10.1111/j.1755-148X.2011.00837.x.
42. Malumbres, M.; Barbacid, M. RAS Oncogenes: The First 30 Years. *Nat Rev Cancer* **2003**, *3* (6), 459–465. https://doi.org/10.1038/nrc1097.
43. Daud, A.; Bastian, B. C. Beyond BRAF in Melanoma. In *Therapeutic Kinase Inhibitors*; Mellinghoff, I. K., Sawyers, C. L., Eds.; Current Topics in Microbiology and Immunology; Springer Berlin Heidelberg: Berlin, Heidelberg, 2010; Vol. 355, pp 99–117. https://doi.org/10.1007/82_2011_163.
44. Fedorenko, I. V.; Gibney, G. T.; Smalley, K. S. M. NRAS Mutant Melanoma: Biological Behavior and Future Strategies for Therapeutic Management. *Oncogene* **2013**, *32* (25), 3009–3018. https://doi.org/10.1038/onc.2012.453.
45. Curtin, J. A.; Fridlyand, J.; Kageshita, T.; Patel, H. N.; Busam, K. J.; Kutzner, H.; Cho, K.-H.; Aiba, S.; Bröcker, E.-B.; LeBoit, P. E.; Pinkel, D.; Bastian, B. C. Distinct Sets of Genetic Alterations in Melanoma. *N Engl J Med* **2005**, *353* (20), 2135–2147. https://doi.org/10.1056/NEJMoa050092.
46. Jakob, J. A.; Bassett, R. L.; Ng, C. S.; Curry, J. L.; Joseph, R. W.; Alvarado, G. C.; Rohlfs, M. L.; Richard, J.; Gershenwald, J. E.; Kim, K. B.; Lazar, A. J.; Hwu, P.; Davies, M. A. NRAS Mutation Status Is an Independent Prognostic Factor in Metastatic Melanoma. *Cancer* **2012**, *118* (16), 4014–4023. https://doi.org/10.1002/cncr.26724.

47. Lee, J.-H.; Choi, J.-W.; Kim, Y.-S. Frequencies of BRAF and NRAS Mutations Are Different in Histological Types and Sites of Origin of Cutaneous Melanoma: A Meta-Analysis: BRAF and NRAS Mutations in Melanoma. *British Journal of Dermatology* **2011**, *164* (4), 776–784. https://doi.org/10.1111/j.1365-2133.2010.10185.x.
48. Thumar, J.; Shahbazian, D.; Aziz, S. A.; Jilaveanu, L. B.; Kluger, H. M. MEK Targeting in N-RAS Mutated Metastatic Melanoma. *Mol Cancer* **2014**, *13* (1), 45. https://doi.org/10.1186/1476-4598-13-45.
49. Devitt, B.; Liu, W.; Salemi, R.; Wolfe, R.; Kelly, J.; Tzen, C.; Dobrovic, A.; McArthur, G. Clinical Outcome and Pathological Features Associated with NRAS Mutation in Cutaneous Melanoma. *Pigment Cell Melanoma Res* **2011**, *24* (4), 666–672. https://doi.org/10.1111/j.1755-148X.2011.00873.x.
50. Dinter, L.; Karitzky, P. C.; Schulz, A.; Wurm, A. A.; Mehnert, M.; Sergon, M.; Tunger, A.; Lesche, M.; Wehner, R.; Müller, A.; Käubler, T.; Niessner, H.; Dahl, A.; Beissert, S.; Schmitz, M.; Meier, F.; Seliger, B.; Westphal, D. BRAF and MEK Inhibitor Combinations Induce Potent Molecular and Immunological Effects in NRAS mutant Melanoma Cells: Insights into Mode of Action and Resistance Mechanisms. *Intl Journal of Cancer* **2024**, *154* (6), 1057–1072. https://doi.org/10.1002/ijc.34807.
51. Valdez-Salazar, F.; Jiménez-Del Río, L. A.; Padilla-Gutiérrez, J. R.; Valle, Y.; Muñoz-Valle, J. F.; Valdés-Alvarado, E. Advances in Melanoma: From Genetic Insights to Therapeutic Innovations. *Biomedicines* **2024**, *12* (8), 1851. https://doi.org/10.3390/biomedicines12081851.
52. Park, J. H.; Pyun, W. Y.; Park, H. W. Cancer Metabolism: Phenotype, Signaling and Therapeutic Targets. *Cells* **2020**, *9* (10), 2308. https://doi.org/10.3390/cells9102308.
53. Sarkar, D.; Leung, E. Y.; Baguley, B. C.; Finlay, G. J.; Askarian-Amiri, M. E. Epigenetic Regulation in Human Melanoma: Past and Future. *Epigenetics* **2015**, *10* (2), 103–121. https://doi.org/10.1080/15592294.2014.1003746.
54. Karami Fath, M.; Azargoonjahromi, A.; Soofi, A.; Almasi, F.; Hosseinzadeh, S.; Khalili, S.; Sheikhi, K.; Ferdousmakan, S.; Owrangi, S.; Fahimi, M.; Zalpoor, H.; Nabi Afjadi, M.; Payandeh, Z.; Pourzardosht, N. Current Understanding of Epigenetics Role in Melanoma Treatment and Resistance. *Cancer Cell Int* **2022**, *22* (1), 313. https://doi.org/10.1186/s12935-022-02738-0.
55. Poulikakos, P. I.; Sullivan, R. J.; Yaeger, R. Molecular Pathways and Mechanisms of BRAF in Cancer Therapy. *Clinical Cancer Research* **2022**, *28* (21), 4618–4628. https://doi.org/10.1158/1078-0432.CCR-21-2138.
56. Nazarian, R.; Shi, H.; Wang, Q.; Kong, X.; Koya, R. C.; Lee, H.; Chen, Z.; Lee, M.-K.; Attar, N.; Sazegar, H.; Chodon, T.; Nelson, S. F.; McArthur, G.; Sosman, J. A.; Ribas, A.; Lo, R. S. Melanomas Acquire Resistance to B-RAF(V600E) Inhibition by RTK or N-RAS Upregulation. *Nature* **2010**, *468* (7326), 973–977. https://doi.org/10.1038/nature09626.
57. Shen, T.; Gao, K.; Miao, Y.; Hu, Z. Exogenous Growth Factors Enhance the Expression of Cola1, Cola3, and Elastin in Fibroblasts via Activating MAPK Signaling Pathway. *Mol Cell Biochem* **2018**, *442* (1–2), 203–210. https://doi.org/10.1007/s11010-017-3204-9.
58. Tan, J.; Zhou, Z.; Feng, H.; Xing, J.; Niu, Y.; Deng, Z. Data-Independent Acquisition-Based Proteome and Phosphoproteome Profiling Reveals Early Protein Phosphorylation and Dephosphorylation Events in Arabidopsis Seedlings upon Cold Exposure. *IJMS* **2021**, *22* (23), 12856. https://doi.org/10.3390/ijms222312856.
59. Fernandez, M. F.; Choi, J.; Sosman, J. New Approaches to Targeted Therapy in Melanoma. *Cancers* **2023**, *15* (12), 3224. https://doi.org/10.3390/cancers15123224.
60. Li, C.; Kuai, L.; Cui, R.; Miao, X. Melanogenesis and the Targeted Therapy of Melanoma. *Biomolecules* **2022**, *12* (12), 1874. https://doi.org/10.3390/biom12121874.
61. Randic, T.; Kozar, I.; Margue, C.; Utikal, J.; Kreis, S. NRAS Mutant Melanoma: Towards Better Therapies. *Cancer Treatment Reviews* **2021**, *99*, 102238. https://doi.org/10.1016/j.ctrv.2021.102238.

Authors

Aisha Torgautova is an A-level student at Crimson Global Academy, located in Almaty, Kazakhstan. Her interest in science and medicine began in her early childhood and has been growing steadily since then. She is eager to further her development in the direction of biomedical and medical research.

IJHSR ijhighschoolresearch.org

RESEARCH ARTICLE

Bridging the Gap: How Charging Infrastructure Shapes EV Adoption in Rural America

Bárbara H. Vargas

PrepaTec, Hacienda Centenario 533, Col. Puerta de Hierro, Monterrey, Nuevo León, 64346, México; barbie.122008@gmail.com
Mentor: Dr. Bimpli

ABSTRACT: Abstract — Sustainable electrification of mobility stems from electric vehicles (EVs), which unfortunately have the lowest penetration in the U.S. rural areas. Unavailability of charging infrastructure proves to be the biggest constraint for adoption. This study looks forward to a deeper understanding of the specific local infrastructure requirements, the socioeconomic context of the regions, and different strategies for investment that need to be put in place to accelerate the adoption of EVs in rural areas. Using data collected from the U.S. Department of Energy's Alternative Fuels Data Center (AFDC), EV market reports, and census data, this research examines the minimum infrastructure density required to sustain adoption, the lags in infrastructure backup across regions, and the logic for investment clusters. Moreover, rural case studies undertaken demonstrate why there are disparate rates of success in adoption, and what challenges and possibilities exist in rural areas. The research impacts EV transition by providing strategic insights into the role of government and other interested parties, and in particular, how investment in short-term charging networks will translate into long-term health benefits to rural places and the broader EV market.

KEYWORDS: Earth and Environmental Sciences, Environmental Engineering, Environmental Effects on Ecosystems, Pollution Control.

Introduction

The world of transportation is gradually turning electric with the rise of electric vehicles (EVs). However, many rural regions in the United States still feel disconnected from this new and contemporary way of transportation. The main reason appears to be the lack of charging stations present in rural regions. It is seen that the availability of charging stations is a crucial aspect of customer growth and confidence. But lower public funding, greater travel distance, and fewer charging options lead to rural areas facing distinct problems. Moreover, although numerous researchers have noted the growth of EV adoption in urban regions, a significant gap remains in the development of rural areas and the rest of the country as a whole. This variation raises an important question: To what extent does charging infrastructure affect EV adoption in rural communities?[1]

Many studies have agreed that charging infrastructure does play a significant role in the adoption of EVs. One example is Sierzchula, who discovered that having more public charging stations per person was more important than financial incentives in 30 different countries. Moreover, in the United States, newer studies have demonstrated that both public and private infrastructure help increase adoption, although private chargers may have a bigger impact, around 16% more EVs for each extra percentage point of private coverage. Also, charging networks tend to divide into two phases: first, the more charging stations you build, the more people feel confident to buy EVs; then, as more people start to buy EVs, the demand becomes higher, and so it requires more charging stations. This process is often known as an indirect network effect, and it is especially important in rural areas due to some rural regions not having any public charging stations within 25 square miles, while some urban areas have over 500.

Still, it is not about how many charging stations there are, but other factors like the type of charger, income levels, education, and even regional culture can affect the adoption of EVs. One study by Khan explains that infrastructure is not disturbed equally; richer and more urbanized areas tend to have better access, which helps increase the inequality in who gets to benefit from the EVs.[2]

There's a gap in the infrastructure when it comes to EV charging stations in rural American areas. Not only is it necessary for a new infrastructure-minded approach, but more focus needs to be brought to the socioeconomic factors to resolve said issues. In doing so, these measures would ensure sustainable transportation across every region.[3]

Oftentimes, urban regions have a volume of innovations, new developments, and investments. This direct shift of focus towards the city puts rural areas at a loss. Recent investigations have noted that the locations of charging stations directly impact the electric vehicle (EV) market in rural America. This paper acknowledges a key limitation, which is the possibility of reverse causality, meaning that while it may appear that charging infrastructure leads to higher EV adoption, it is equally possible that higher EV demands attract infrastructure investments. Despite the fact that this study does not focus on methods to resolve this causality, it does look to provide insights by analyzing patterns, socioeconomic indicators, and policy differences between these two regions.[4]

This research paper is based on two main ideas. The indirect network effect, which is when people see that there are enough chargers around, they feel less worried about running

DOI: 10.36838/v8i2.103

out of battery power. That makes them more likely to buy an EV. Then, as more people switch, it becomes worthwhile to build more stations, which repeats the cycle. This is a key point in rural areas where infrastructure is limited. Then the second main idea is socioeconomic access. Infrastructure needs to be studied in relation to people's income, education, and even how far they live from a charging station. Now, if we only count the number of stations without thinking about who has access to them, we might be missing important factors that affect the results. Moreover, I have included some variables like rural and urban locations, and some income levels, as suggested in the recent study by Khan.

■ Methods

This research relied on secondary information from several reliable external sources to assess EV adoption patterns and charging infrastructure expansion. The U.S. Department of Energy's Alternative Fuels Data Center (AFDC) supplied extensive information sources across the nation, whether public or private, which allowed spatial analysis of infrastructure coverage. Further, EV Market Reports were referenced to gain insight into sales patterns, behavioral changes in adoption, and expected growth in the EV market, including regional adoption trends. Additionally, the use of Census Data enabled the evaluation of demographic and socioeconomic parameters of income levels and population concentration in less populated areas to interpret adoption patterns.

Now, moving on to the Geospatial and Statistical Analysis. To understand how charging infrastructure is distributed, GIS geo-visualization software was utilized to map EV charger locations by area. This visual approach helped identify infrastructure gaps, especially in rural regions, and allowed for clearer comparisons between areas of EV adoption. In addition to this information, a correlation analysis was conducted to establish the relationship between EV sales per capita and the density of charging stations. This helped us determine whether areas with more infrastructure saw proportionally higher adoption rates. Additionally, to build on these findings, a multiple linear regression model was developed to estimate the impact of infrastructure and income levels on EV adoption.[5]

The model incorporated key variables such as the number of chargers per square mile, median household income, and a rural versus urban classification. Commands were included for population size and regional differences to strengthen the results.

Moreover, to explore spatial disparities more concisely, this study conducted a regional analysis that classifies countries into three categories—low, medium, or high—in terms of charging infrastructure density. Within these categories, EV adoption rates were compared and analyzed alongside demographic data to identify any barriers or regional constraints.[6]

Furthermore, a rural versus urban comparison was carried out within individual states, particularly focusing on two case studies, which are rural Kentucky and urban San Francisco. This approach helped control for state-level policy and funding environment while revealing localized inequalities in access and adoption.

■ Result and Discussion

3.1 Data Analysis and Statistical Findings:

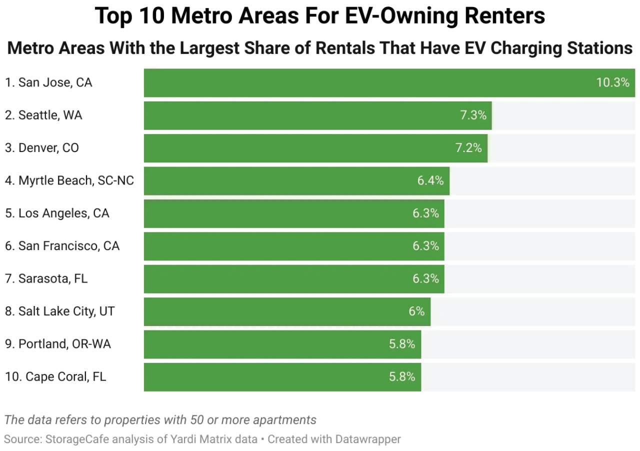
The analysis demonstrated significant variations in electric vehicle adoption rates between urban and rural areas in the United States, specifically between the San Francisco Bay Area and the state of Kentucky. The correlation analysis demonstrated a strong positive relationship between the availability of EV charging infrastructure and adoption rates, with a correlation coefficient of $r = 0.78$ and $p < 0.01$. This indicates that areas with higher concentrations of charging stations tend to see faster growth in EV usage, confirming the importance of infrastructure availability in enabling adoption.[7]

For a better understanding of this relationship, a linear regression model was applied, incorporating charging station density, median income, and even education levels as independent variables. The model accounted for 68% of the variation in EV adoption growth $R^2 = 0.68$ Basically, in urban areas like the San Francisco Bay Area, charging station density rose as the most influential factor with a standardized coefficient of $\beta = 0.56$ and $p < 0.01$. While in rural areas such as Kentucky, infrastructure still showed a statistically significant effect, although weaker, with $\beta = 0.32$ and $p < 0.05$. These results emphasize the regional difference not only in infrastructure but also in how effectively it adapts to adoption.[8]

CAUTION: The reliability of data from rural areas is limited due to underreporting and the lack of disaggregated EV sales data.

3.2 Figures, Graphs, and Equations:

Figure 1. The top 10 largest metro areas for electric vehicle charging stations.[9]


Top 10 Metro Areas For EV-Owning Renters

Metro Areas With the Largest Share of Rentals That Have EV Charging Stations

The data refers to properties with 50 or more apartments

Source: StorageCafe analysis of Yardi Matrix data • Created with Datawrapper

Figure 1: Shows the top 10 metro areas in the U.S. where renters have the best access to EV charging stations. San Jose is at the top with 10.3%, and San Francisco comes in sixth with 6.3%. This supports the idea that better access to chargers, especially in urban areas, really helps boost EV adoption.

The figure from above, Figure 1, shows the top ten metropolitan areas in the United States for electric vehicle owning renters. Within these areas, San Francisco ranks sixth place with 6.3% of rentals offering access to charging infrastructure. This supports the claim that infrastructure accessibility plays an important role in regional EV adoption.

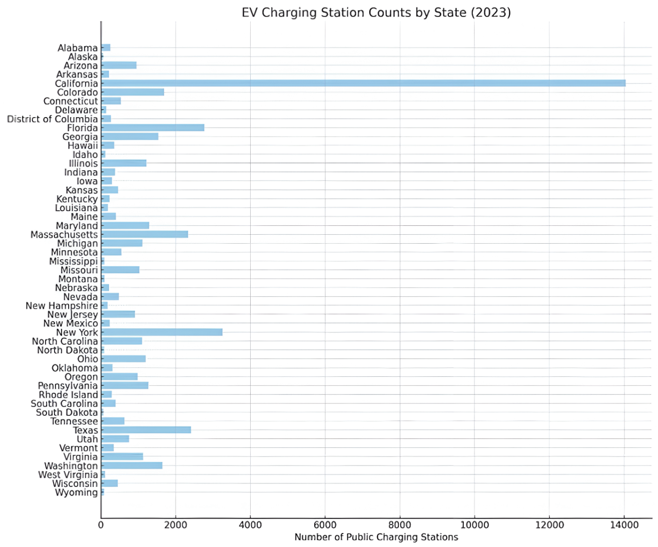


Figure 2: The graph shows how many public charging stations each U.S. state had in 2023. California is way ahead with 14,000 stations, while states like Kentucky have many fewer. This big difference helps explain why rural areas are struggling more with EV adoption compared to cities.

As mentioned before, Figure 2 presents a comparison of public EV charging stations by state as of 2023. California surpasses other states, with over 14,000 stations, while Kentucky has significantly fewer. The visual contrast between these two states illustrates the infrastructure gap that contributes to adoption inequality between urban and rural contexts.

The regression equation used in the analysis is

Equation 1: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

where Y represents the annual growth rate in EV adoption, X_1 is the charging station density (measured in stations per 1,000 square miles), X_2 is the median household income, and ε is the error term. Now, in urban regions, the coefficient for charging station density was 0.56 and statistically significant at $p < 0.01$, while income had a slightly lower influence with = 0.41 and $p < 0.05$. In rural regions, the effect of the infrastructure was still significant with $\beta = 0.32$; however, the influence of income was more muted. This demonstrates accurately and confirms that the infrastructure plays an essential role in both settings, although its effect is amplified in urban areas.11

3.3 Reflections and Limitations:

The study faced several limitations that must be mentioned. First, data on private charging stations were not included in the analysis, which may have resulted in an underestimation of the actual infrastructure availability, particularly in urban settings where home-based charging is more common. Second, the lack of unfiltered rural data imposed some challenges, as EV sales figures were often aggregated at the state level, making it even more difficult to isolate patterns within smaller rural communities. Third, the regression model did not incorporate variables such as cultural resistance to electric vehicles, grid capacity constraints, or the presence of state-level incentives, which may have influenced adoption outcomes, especially in rural areas. Lastly, the dataset simplified demographics by focusing mainly on income and education, potentially overlooking other relevant factors such as household size, previous vehicle ownership habits, or even awareness of environmental policy.[12]

Despite these limitations, the consistency of our findings across different statistical methods strengthens the credibility of the study's conclusions.

3.4 Interpretation of Findings:

The main hypothesis of the study, as mentioned before, that urban areas such as the San Francisco Bay Area would show significantly higher rates of EV adoption than rural regions like Kentucky, conditional on infrastructure presence and economic factors, was validated by the data. Some urban residents benefit from dense charging networks, shorter average travel distances, and even higher income levels, all of which lower practical and psychological barriers to EV usage.[13]

The data also suggest that modest infrastructure in rural areas is not enough to stimulate widespread adoption. While EV chargers are indeed present in states like Kentucky, the low adoption rate indicates that other unmeasured variables, such as cultural hesitancy or lack of policy outreach, may be playing a role.

■ Conclusion

This paper analyses the role of charging infrastructure and household income on the rates of EV uptake using the San Francisco Bay Area and Kentucky as representative case studies for urban and rural territories, respectively. The results indeed confirmed the hypothesis, with regard to both charging station density and household income: these factors appear to be influencing EV adoption significantly. The urban centers presented more infrastructure and population that facilitated higher adoption rates, while the rural areas suffered from inadequate infrastructure and poor economies, which made the uptake rate quite dismal.[14]

Even though this paper provides new perspectives that reinforce the importance of physical infrastructure and economic capacity, it raises additional questions for future research consideration. Why do areas with medium charging infrastructure density exhibit slower adoption rates? What role do cultural attitudes and psychological factors play in influencing EV adoption, particularly in rural communities? Future research should focus on these aspects, incorporating qualitative data from EV adopters and non-adopters to better understand the underlying barriers. Additionally, exploring the impact of specific policy measures, outreach programs, and emerging technologies like portable chargers could provide valuable insights into overcoming rural infrastructure deficits.

This research highlights the importance of adopting proper approaches in rural areas while there is still an ongoing urban adoption race. Significant allocation of resources into rural charging infrastructure and subsidies directed at income-generating activities, as well as setting up region-specific strategies, are essential to enable a smoother movement to sustainable means of transportation. Equal access and popularity of electric mobility to targeted rural locations, coupled with creative and joint efforts, will help to bridge the urban-rural EV adoption gap.

Acknowledgments

For this paper, I would like to acknowledge my parents for all the support, Dr. Bimpli, who gave me mentorship and tips to make this possible, and Daryn Howland, who was my TA, for giving me feedback on my paper draft.

References

1. JSmith. Charging Ahead: Rural Communities and the Transition to Zero-Emission Vehicles - Community Energy Association. https://www.communityenergy.ca/rural-communities-zero-emission-vehicles-transition/ (accessed 2025-06-10).
2. IEEE Xplore Full-Text PDF. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9028778 (accessed 2025-06-10).
3. Fox-Penner, P.; Hao, J. (Roger); Hatch, J.; Helveston, et al. The Critical Role of Public Charging Infrastructure. https://open.bu.edu/bitstream/2144/39112/1/MeltingIceBook_Web_FINAL.pdf (accessed 2025-06-10).
4. Shevchenko, E. THE ROLE OF CHARGING INFRASTRUCTURE AND INCENTIVES ON THE ADOPTION OF ELECTRICAL VEHICLES IN THE UNITED STATES. University of the Incarnate Word. https://edeconomy.com/wp-content/uploads/2023/08/THE-ROLE-OF-CHARGING-INFRASTRUCTURE-AND-INCENTIVES-ON-THE-ADOPTION-OF-ELECTRICAL-VEHICLES-IN-THE-UNITED-STATES.pdf (accessed 2025-06-10).
5. Charging Summit, E. The State of EV Charging Infrastructure in Rural Areas - EV Charging Summit Blog. https://evchargingsummit.com/blog/the-state-of-ev-charging-infrastructure-in-rural-areas/ (accessed 2025-06-10).
6. Bevis, K.; Smyth, A.; Walsh, S. Plugging the Gap – Can Planned Infrastructure Address Resistance to Adoption of Electric Vehicles? https://core.ac.uk/download/29840396.pdf (accessed 2025-06-10).
7. Singh, V. How Can California Best Promote Electric Vehicle Adoption? The Effect of Public Charging Station Availability on EV Adoption. https://core.ac.uk/download/215289196.pdf (accessed 2025-06-10).
8. View of the Influence of Electric Vehicle Charging Stations on the Quality of Power Supply to the Consumer in Novosibirsk 10/0.4 KV Electrical Networks. https://esrj.ru/index.php/esr/article/view/243/198 (accessed 2025-06-10).
9. DiveImage. Map Displaying Top 10 Metro Areas for EV Adoption; Dive, 2023. https://imgproxy.divecdn.com/UkqB4rKCp_fUJboWK7xlrQugMmRjALDERSvwU57UFFM/g:ce/rs:fill:1220:862:0/Z3M6Ly9kaXZlc2l0ZS1zdG9yYWdlL2RpdmVpbWFnZS8tc3Bhbi1zdHlsZS1kaXNwbGF5LWJsb2NrLXdpZHRoLTEwMC10ZXh0LWFsaWduLWNlbnRlci10b3AtMTAtbWV0cm8tYXJlYXMtZm9yLWVfZWZzU1lCSC5wbmc=.webp (accessed 2025-06-10).
10. InsideEVs. EV Charging Stations by State, 2023; Shopify, 2023. https://cdn.shopify.com/s/files/1/2948/6296/files/ev_charging_stations_by_state_2023_600x600.png?v=1728956702 (accessed 2025-06-10).
11. Fox, S. J. Planning for Density in a Driverless World. https://core.ac.uk/download/70375689.pdf (accessed 2025-06-10).
12. U.S. Electric Vehicle Charging Infrastructure Market Size, Share & Trends Analysis Report by Charger Type (Slow Charger, Fast Charger), by Connector Type, by Level of Charging, by Connectivity, by Application, and Segment Forecasts, 2025 - 2030. https://www.grandviewresearch.com/industry-analysis/us-electric-vehicle-rging-infrastructure-evci-market# (accessed 2025-06-10).
13. Manansala, J. Number of EV Charging Stations by State: 2024 Overview. Lectron EV, 2024. https://ev-lectron.com/blogs/blog/number-of-ev-charging-stations-by-state-2024-overview (accessed 2025-06-10).
14. Mark Gbeda, J. Effects of Charging Infrastructure on Electric Vehicles Adoption and Transportation Emissions in the United States: A Panel Analysis. https://dr.lib.iastate.edu/server/api/core/bitstreams/f95f010a-fe78-4cc1-807c-2c417b9b4610/content (accessed 2025-06-10).

Author

Bárbara Hernández Vargas is a high school student at PrepaTec Cumbres in Monterrey, Mexico. She is passionate about sustainability, electric mobility, and innovation. She plans to pursue Mechanical or Aeronautical Engineering at Johns Hopkins or MIT universities and recently participated in the Johns Hopkins Engineering Innovation Program.

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