INTERNATIONAL JOURNAL OF PHARMACEUTICAL SCIENCES
[ISSN: 0975-4725; CODEN(USA): IJPS00]
Journal Homepage: https://www.ijpsjournal.com

OPEN ACCESS

Review Article

A Step Towards Preventive Care: Leveraging Machine Learning for Heart Disease Risk Prediction

Muskan Khepar, Ashish Sharma

Punjab Engineering College, Chandigarh

ARTICLE INFO

Published: 11 Oct. 2024
Keywords:
Machine Learning, Exploratory Analysis, Data Mining, Logistic Regression, Heart Attack Risk Assessment, And Cardiovascular Disease.
DOI:
10.5281/zenodo.13920479

ABSTRACT

Predicting heart disease has been one of the most challenging jobs in the medical industry in recent decades. Heart disease claims the lives of about one person per minute in the modern era. Processing vast amounts of data in the healthcare industry requires the use of data science. Since predicting cardiac illness is a difficult undertaking, it is necessary to automate the process in order to minimize risks and notify patients well in advance. The Kaggle machine learning repository's dataset on heart disease is used in this paper. The suggested work uses logistic regression to categorize patients' risk level and forecast their likelihood of developing heart disease. As a result, this research provides a thorough analysis of the machine learning algorithm's performance in detecting the risk of a heart attack. The testing results confirm that, when compared to other ML algorithms used, the Logistic Regression approach has the greatest accuracy of 90.81%.

INTRODUCTION

Cardiovascular diseases (CVDs) are the leading cause of death worldwide, responsible for 17.9 million deaths annually, or nearly one-third of all global deaths. Over 85% of these are due to heart attacks and strokes. The global CVD mortality rate stands at 233 cases per 100,000 people, with low- and middle-income countries seeing a rapid increase in cases and deaths. Premature CVD deaths are projected to rise to over 23 million annually by 2030. To combat this, health systems need strengthening, and efforts must focus on prevention, early detection, and treatment through evidence-based strategies. The bright side is that heart attacks are remarkably avoidable, and simple lifestyle changes (including cutting back on alcohol and tobacco use, eating healthfully, and exercising) along with prompt treatment greatly increase the likelihood of survival. Despite this, the multifactorial nature of many contributing risk factors, such as diabetes, hypertension, increased cholesterol, and so on, makes it challenging to identify high-risk patients. Information mining and

***Corresponding Author:** Muskan Khepar
Address: *Punjab Engineering College, Chandigarh*
Email ✉: kheparmuskan@gmail.com

Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

machine learning take center stage here. Because machine learning (ML) tactics are more common in design recognition and characterization than other traditional factual methodologies, experts and academics have also turned to ML strategies to construct screening devices.

A. Heart Attack

A coronary artery spasm, commonly referred to as a heart attack, typically happens when a blood clot obstructs the heart's blood supply. Tissue dies when it is oxygen-starved and lacks blood. With heart attacks contributing to almost 25% of all fatalities globally in 2023, they are the leading cause of mortality globally. Anxiety, exhaustion, dizziness, tightness or pain in the arms, back, neck, or chest, and an irregular pulse are common signs of a heart attack. Not every person experiencing a respiratory failure will exhibit the same symptoms or level of severity. Some suffer from mild anguish, while others have more severe agony. Some folks have no symptoms at all. For a few, unexpected heart failure may be the main symptom. However, the greater the number of symptoms and indicators you have, the more likely it is that you are experiencing respiratory failure.

While some heart attacks happen suddenly, many people experience warning signs and symptoms hours, days, or even weeks in advance. The earliest warning signs could be weight or sporadic chest pain that is relieved by rest and triggered by activity (angina). A brief reduction in the blood supply to the heart causes angina.

B. Heart Attack Causes

When obstruction occurs in one or more of your coronary veins, you may experience cardiovascular failure. Gradually, the accumulation of fatty deposits, like as cholesterol, forms deposits known as plaques that can narrow the conduits (atherosclerosis). The majority of cardiovascular failures are caused by a disease known as coronary conduit sickness.

A plaque rupture that releases various chemicals and cholesterol into the bloodstream can occur during a myocardial infarction. At the location of the fracture, a blood coagulation forms. If coagulation is too high, it can obstruct blood flow via the coronary artery, depriving the heart of oxygen and nutrients (ischemia).

The coronary artery may be completely or partially blocked.

- A total blockage suggests that you have experienced a myocardial localized necrosis (STEMI) of the ST height.
- A partial blockage suggests that you have experienced a localized necrosis of the heart that is not ST height (NSTEMI).

Depending on the type you've had, there may be differences in determination and therapy.

Blockages resulting from atherosclerosis do not trigger heart attacks in all cases. In the absence of obstructive coronary artery disease, a heart attack caused by other heart and blood vessel disorders is referred to as a myocardial infarction (MINOCA). MINOCA is more prevalent among women, younger individuals, and members of racial and ethnic minorities, such as Asian, Black, and Hispanic/Latino persons. Different conditions can affect the heart differently and lead to MINOCA.

Although **little plaques** in your arteries may not obstruct your blood vessels, they may fracture or lose part of their outer covering. Blood clots may form on these plaques as a result. Your coronary arteries may then become blocked by the blood clots. Small plaque formation is more prevalent in women, smokers, and those with other blood vessel disorders. Even in the absence of plaque accumulation, a **sudden and severe spasm (tightening) of your coronary artery** might obstruct blood flow through it. One of the risk factors for a coronary spasm is smoking. If you smoke, you can be more susceptible to spasms brought on by extremely cold temperatures or stressful circumstances. Cocaine and other drugs

can also induce myocardial spasm. When a blood clot enters your bloodstream and lodges in your coronary artery, it is known as a **coronary artery embolism**. This may stop your artery's blood flow. This is particularly likely in those with atrial fibrillation or illnesses like thrombocytopenia or pregnancy that increase the risk of blood clots.

When a tear develops inside your coronary artery, it can lead to **spontaneous coronary artery dissection, or SCAD**. The ripped tissue may subsequently obstruct your artery or a blood clot may form at the site of the rupture. Pregnancy, high levels of physical activity, and stress can all contribute to SCAD. Marfan syndrome sufferers and women under 50 years old who are pregnant are more likely to experience this illness.

C. Risk Factors

The likelihood of developing coronary artery disease and experiencing a heart attack is increased by specific risk factors.

Factors under control include:

Lifestyle choices include things like smoking, not getting enough exercise, eating a poor diet that includes a lot of foods high in sodium or saturated fat.

Additional medical disorders include high blood pressure, high blood sugar, diabetes, high blood triglycerides, high blood cholesterol, and preeclampsia (high blood pressure during pregnancy), being overweight or obese.

It is referred to as metabolic syndrome if a person possesses three or more of these heart disease-related risk factors. This significantly raises the chance of having a heart attack.

Factors not under control include:

Age: Men are more likely to develop heart disease after 45, and women are more likely to do so after 55 (or after menopause).

Early heart disease in the family: If your mother, sister, or father were diagnosed with coronary artery disease before the age of 65, or if your brother or father was diagnosed with the condition before the age of 55, you are at an increased risk of developing early heart disease viral and bacterial infections

Prevention:

By treating any known coronary artery disease or altering behaviors that increase your risk of heart attack, you can reduce your chance of having one. Heart disease can be prevented by adopting heart-healthy lifestyle practices, such as eating a balanced diet, exercising regularly, giving up smoking, controlling stress, and keeping a healthy weight. These modifications can reduce the risk of a heart attack even in those with pre-existing coronary artery disease.

D. Difficulties

The damage your heart sustains after a respiratory failure is often linked to complications, which might lead to:

Arrhythmias, or irregular heartbeats: Electrical "short circuits" can result in irregular heartbeats, some of which can be fatal and cause death.

Heart disappointment: You may have so much damage to your heart from a respiratory failure that your heart's remaining muscle is unable to pump enough blood out of your heart. Cardiovascular breakdown can occur suddenly or develop gradually because of widespread, lifelong heart damage.

Abrupt heart failure: An electrically unsettling influence that creates an irregular heartbeat (arrhythmia) causes your heart to quit beating suddenly. Acute heart failure increases the risk of cardiac failure, which can result in death without immediate medical attention.

E. Problem Statement

The leading cause of illness and death worldwide is heart disease:

- Compared to other causes, it is the reason behind more deaths each year. Thus, there is great potential for the development and usefulness of illness prevention and detection for most cases.

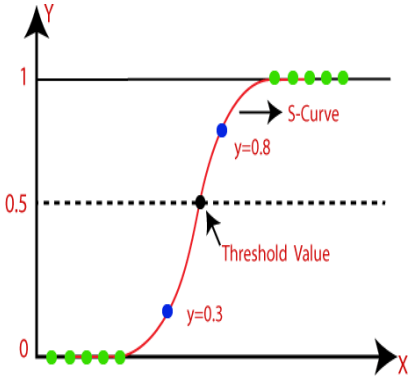
- Nevertheless, due to the multifactorial character of many contributing risk factors, including diabetes, high blood pressure, high cholesterol, and so forth, it is challenging to identify people who are at high risk.

Here's when data mining methods and machine learning algorithms come in handy.

PROPOSED EXPERIMENTAL METHOD

A. Logistic Regression

An approach for supervised learning classification that estimates the likelihood of a target variable is called logistic regression. There would only be two classes conceivable due to the dichotomous character of the dependent or target variable.



An elucidation of the standard logistic function can precede an explanation of logistic regression. Any real number between zero and one can be entered into the logistic function, which is a sigmoid function. It is described as

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$

The following are some major benefits of employing logistic regression:

- Lowers the danger of overfitting
- Does not make any assumptions about the distributions of classes in feature space
- Is simpler to construct, analyze, and train

B. Confusion Matrix

A confusion matrix, commonly called an error matrix, is a type of standard table structure used to visualize algorithm performance, especially when dealing with statistical classification difficulties. A supervised learning one is sometimes called a matching matrix in unexpected learning. In the matrix, a hypothetical class is represented by each row, while instances in the real class are represented by each column (or vice versa). The name alludes to how easy it is to ascertain whether the two classes are being confused by the system, or if they are typically confused with one another. That's a specific type of contingency table where the two dimensions (a variable is a combination of each class and dimension in the dimension table) have the same set of "classes" in both "real" and "approximate" dimensions.

		Prediction	
		0	1
Actual	0	TN	FP
	1	FN	TP

- **False positive (FP):** A test result that falsely indicates the existence of a particular condition or characteristic
- **True positive (TP):** Also referred to as true positive rate or likelihood of detection in many fields, true positive (TP) counts the percentage of true positivity that is successfully identified.
- **True Negative (TN):** Also referred to as Specificity (or actual negative rate), True Negative (TN) gauges the percentage of true negatives that are reliably identified.
- **False Negative (FN):** Tests that get a FN result indicate a condition does not exist while it exists. As an illustration, consider a test result that indicates a person does not have cancer while they do.

Until a call threshold is considered, logistic regression does not become a classification method. Setting the threshold value is an important logistic regression component that is dependent on the classification problem.

The evaluation of "the price | the worth" of the threshold value is adversely affected by the exactness and recall values. While it would be perfect if every exactness and recall were the same,

this rarely happens. We take into account the following elements to establish the threshold in the case of a Precision-Recall trade-off: -

Low Precision/High Recall: We choose a call price that has a low exactitude accuracy or a high recall value when we want to reduce the number of false negatives without lowering the number of false positives. For instance, in a cancer diagnostic application, we frequently classify any affected patient as unaffected without considering whether the patient has received a de jure cancer diagnosis. This may be the case since cancer can be detected in other medical problems, but not in a candidate who has already been rejected.

High Precision/Low Recall: We frequently select a call with a high precision value or a low recall value when we wish to lower the number of false positives without also lowering the number of false negatives. If, for example, we are inclined to classify customers according to how they will respond to a made-to-order advertisement, we want to be 100% sure that the customer will respond favorably to the promotion because, should the customer respond unfavorably, the consumer might miss out on a possible sale.

C. Methodology

The following steps are involved in predicting the risk of a heart attack:

Split the data into test and trainsets

- Import the csv file and necessary libraries
- Clean and preprocess the data before using it
- Determine the variable association analysis

Using the data, train the logistic regression model to:

- Predict test set outcomes
- Compare expected and actual results
- Evaluate findings and make inferences

Importing: Bring in the necessary libraries and the CSV file. The required libraries are as follows:

- NumPy: used for manipulating matrices, arrays, and linear algebra
- Pandas: Used for time series and table data analysis and manipulation
- Sklearn: Offers a range of methods for clustering, regression, and classification.
- Matplotlib: This program plots graphs and shows how different features affect model prediction.

Pre-processing and cleaning of data: Pre-processing refers to the modifications made to our data before it is sent into the algorithm. One technique used to transform the raw data into a clean informational index is called data preprocessing.

The actions needed to prepare the data are as follows:

- Verifying the data set's Null values
- Changing null values to the column's average value
- Dividing the columns into continuous and categorical values
- Making the data set normal

The following are potential obstacles in this step:

- Locating a sizable dataset
- Data formatting
- Data cleaning
- Data sampling
- Augmenting the data.

RESULTS OF EXPERIMENTS AND DISCUSSION

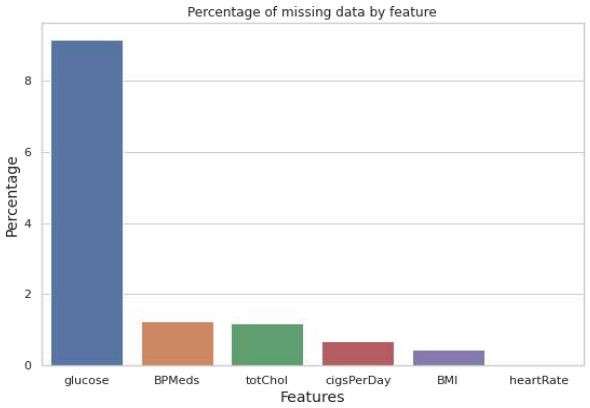
A Kaggle dataset of 3,700 rows (datapoints) and 13 columns (attributes) was utilized. 165 (54%) of the 1,998 datapoints had Cardio Vascular Disease (CVD), while the remaining 1,702 (46%) did not.

A selection from the first three rows of the data

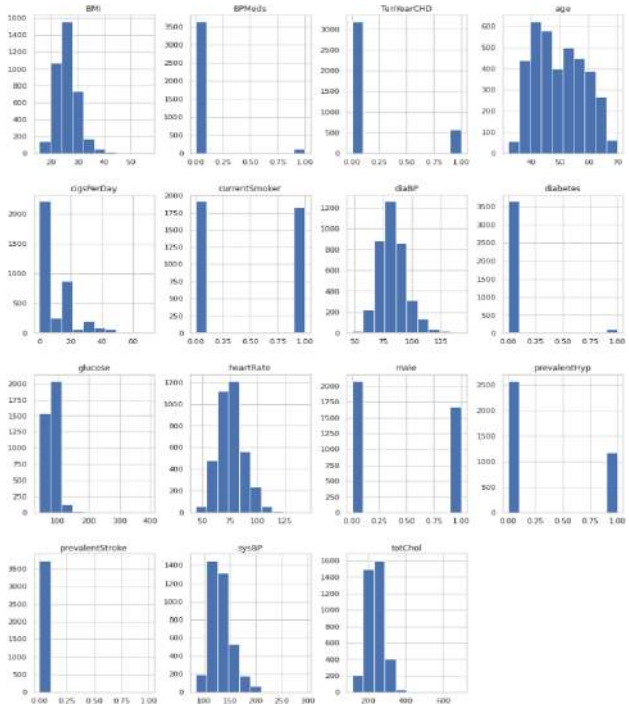
	Id	BMI	BPMeds	Age	Diabetes
0	842302	18.43	0.032	35	0.011
1	842517	25.6	0.045	46	0.062
2	843009	30.1	1.02	57	1.11

Exploratory Analysis

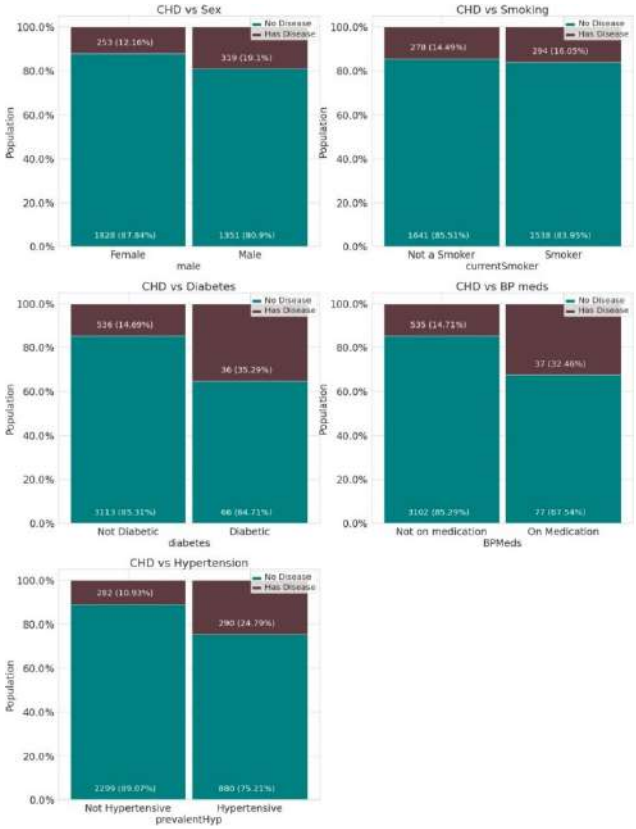
The initial stage involves examining the distribution of various attributes, which is best illustrated using histograms.



Plot the graphs between each column (attribute) and the proportion of patients with and without CVD after that.



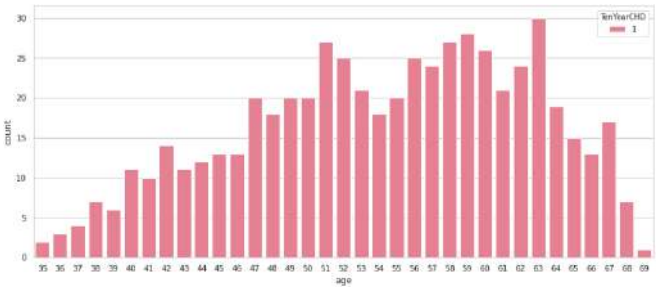
Examine the percentages of good and negative cases in each category to obtain additional understanding.



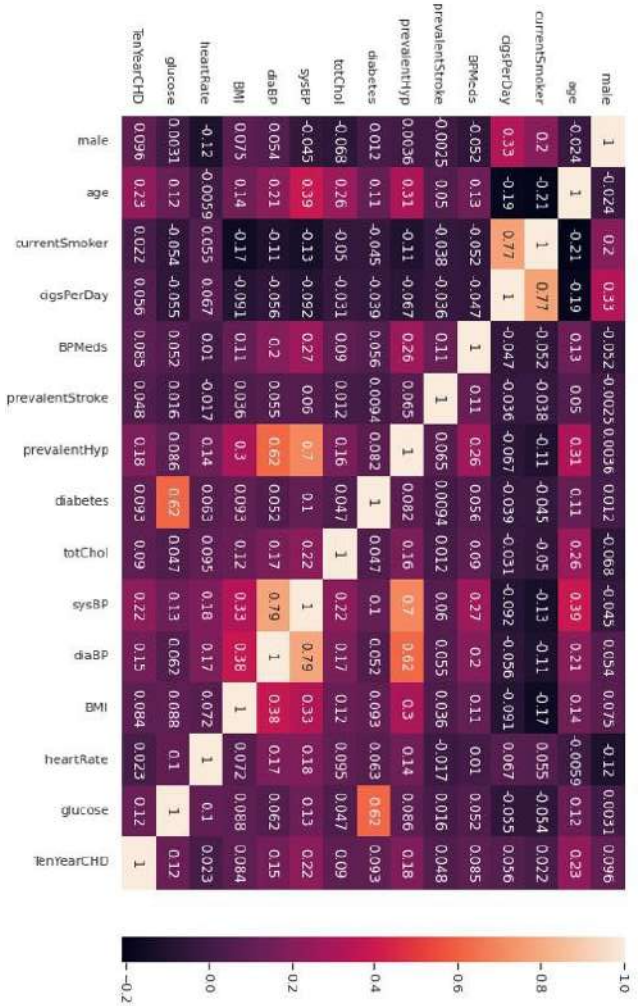
It was challenging to draw conclusions because of the unbalanced nature of the data set, however these are the conclusions that could be inferred from the observations:

- Compared to females, a somewhat higher percentage of men have CHD.
- The proportion of individuals with CHD who smoke and those who do not is nearly equal.
- Compared to individuals without comparable morbidities, a larger percentage of people with CHD are diabetics and those with predominate hypertension.
- A higher proportion of CHD patients take blood pressure medication.

The distribution of CHD patients' ages revealed another intriguing trend: the number of sick people grew progressively older, peaking at 63 years old.



Checking the correlation between the various features and the objective variable, as well as between them, is the last stage. This will disclose any co-linearity among the features and provide a good assessment of how strong the features are as coronary heart disease predictors.



- **Model Results**

The model's accuracy is determined to be 91%.

```
y_pred = logreg.predict(X_test)
print('Accuracy of logistic regression classifier on test set: {:.2f}'.format(logreg.score(X_test, y_test)))
```

```
Accuracy of logistic regression classifier on test set: 0.91
```

The confusion matrix illustration:

```
from sklearn.metrics import confusion_matrix
confusion_matrix = confusion_matrix(y_test, y_pred)
print(confusion_matrix)
```

```
[[1509  195]
 [146   1850]]
```

- True Positives: 1,509
- False Negatives: 195
- False Positives: 146
- True Negatives: 1,850

Classification Report:

```
from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.91	0.89	0.90	35
1	0.90	0.93	0.92	41
accuracy			0.91	76
macro avg	0.91	0.91	0.91	76
weighted avg	0.91	0.91	0.91	76

Other ML algorithms like K-nearest neighbor, Random Forest Classifier, Decision Tree Classifier, Support Vector Machine and XG Boost Machine can also be used to detect the heart attack risk. The dataset was also tested with these algorithms and the performance was compared. The table summarizes and compares the outcomes of different algorithms:

Model	Testing Accuracy%	Precision	F1 score
Logistic Regression	90.81	91	92
K-nearest neighbors	86.81	86	87
Support vector machine	87.91	88	89
Decision tree classifier	78.02	85	82
Random forest classifier	82.42	82	86
XG Boost classifier	83.52	87	87

CONCLUSION

Make the classification or prediction based on the test results and show the patient's actual values—which show whether they have CVD—as well as the classification or prediction the Logistic Regression model made. I have seen that the

model misdiagnosed certain patients as having cardiovascular disease (CVD) when in fact they did not, and it misclassified as cancer-free those who did have CVD. This approach is good, but I want it to be even better in terms of how it affects the lives of other people. It should be as accurate as feasible and perform on par with physicians, if not better. Therefore, there is still some fine-tuning that needs to be done on each model. There are a lot of possible improvements that may be looked into to make this expectation framework more flexible and accurate.

REFERENCES

1. Cristianini, N., and J. Shawe-Taylor. "An Introduction to support vector machines and other kernel-based learning methods" New York: Cambridge University Press, 2000.
2. Vapnik, V. N. "The Nature of Statistical Learning Theory" New York: Springer, 1995.
3. Sonam Nikhar, A.M. Karandikar "Prediction of Heart Disease Using Machine Learning Algorithms" in International Journal of Advanced Engineering, Management and Science (IJAEMS) June 2016 vol-2
4. Deeanna Kelley "Heart Disease: Causes, Prevention, and Current Research" in JCCC Honors Journal
5. Dhafar Hamed, Jwan K. Alwan, Mohamed Ibrahim, Mohammad B. Naeem "The Utilization of Machine Learning Approaches for Medical Data Classification" in Annual Conference on New Trends in Information & Communications Technology Applications – March 2017
6. Theresa Princy R, J. Thomas, Human heart Disease Prediction System using Data Mining Techniques, International Conference on Circuit Power and Computing Technologies, Bangalore,2016
7. Nagaraj M Lutimath, Chethan C, Basavaraj S Pol., Prediction of Heart Disease using Machine Learning, International journal Of Recent Technology and Engineering,8, (2S10), pp 474-477, 2019.

HOW TO CITE: Muskan Khepar, Ashish Sharma, A Step Towards Preventive Care: Leveraging Machine Learning for Heart Disease Risk Prediction, Int. J. of Pharm. Sci., 2024, Vol 2, Issue 10, 566-573. https://doi.org/10.5281/zenodo.13920479