Endothelial Dysfunction Analysis Using Machine Learning

Oviya V[1, a)] and Dr T Jackulin[2, b)]
[1]Student
(Panimalar Engineering College, Chennai -600123).
[2]Associate Professor
(Panimalar Engineering College, Chennai-600123).

Abstract—The term "endothelial dysfunction analysis", which is used to describe a variety of conditions that impact the heart and circulatory system is vast. It is a substantial global contributor to disability. One of the most crucial body parts is the heart, its ailment also affects other body organs. Heart conditions can take many different forms. Heart failure and heart attacks that result in cardiac hypertrophy, coronary artery narrowing or blockage, heart valve dysfunction, and several other conditions are the most popular practices.

Keywords—Machine learning, Endothelial dysfunction, cardiac arrest, heart disease.

INTRODUCTION

In this project, we'll develop a website that analyses endothelial dysfunction and outputs results in accordance with the concerned with the understanding. We use an online dataset that contains specific medical data about patients to assess whether they are more or less likely to have a heart attack. Analyze the dataset using the available data, classify the target variable using a variety of machine learning models, and decide which tactic is most accurate for this dataset. We use that algorithm on our website to identify whether the new data is likely to contain heart disease or not. A spectrum of disorders that harm your heart are referred to as endothelial dysfunction. Blood vessel problems like coronary artery disease are among the heart-related illnesses. 17.9 million individuals worldwide succumb to cardiovascular disease each year, according to WHO stats. This heart ailment is mostly brought on by human living, claims the medical study. There are a number of other significant indicators that the person may or may not be at risk for heart disease in addition to this. Both the health organisation and the patients would greatly benefit if proper machine learning techniques could more accurately forecast heart disease using a dataset of individuals who undergone health examinations.

According to the medical study, this heart condition is primarily caused by human lifestyle. In addition to this, there are a number of important indicators that the person may or may not be at risk for heart disease. If we develop appropriate machine learning methods that more reliably anticipate heart disease using data from people who had health examinations, it is tremendously beneficial to both the health organisation and the patients.

Important information acquired by WHO (World Health Organizations)

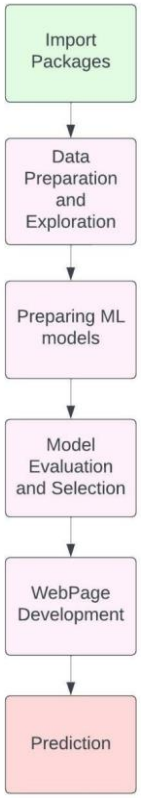
- The primary cause of death in the world is cardio/endothelial disease.
- According to predictions, CVDs were responsible for 17.9 million deaths worldwide in 2019—32% of all fatalities. 85% of these deaths were caused by heart attacks and strokes.
- More than 75% of endothelial dysfunction deaths begin in earnest in developing and intermediate countries.
- In 2019, non-communicable disorders caused 17 million premature deaths (fatalities before the age of 70), and 38% of those deaths were attributed to CVDs.
- Cardiovascular disease must be detected early in order to start treatment with counselling and medicine. The majority of cardiovascular illnesses can be prevented by treating behaviours, such as nicotine, unhealthy eating and obesity, inactivity, and excessive alcohol use.

LITERATURE SURVEY

Endothelial Dysfunction Analysis is one of the most common causes of mortality globally, as per [1]. Medical professionals find it tough to forecast because it is a complex task that calls both experience and advanced knowledge. Nowadays, data mining and machine learning-based medical supporting technologies have an essential function in the prediction of cardio vascular illnesses.[2] As stated by the Centers for Disease Control and Prevention, heart illnesses are the primary cause of death in the nation for men, women, and individuals of the majority of racial and ethnic groups .It results in more than one death per minute and approximately 500,000 deaths yearly, costing billions of dollars. It has been possible to predict if people have cardiac disease using previous machine learning techniques.[3]Analysis and Prediction of Heart Disease Incidence Chala Beyene et al. recommended using data mining approaches. The main objective is to predict the onset of heart illness in order to reliably and rapidly identify the condition. In hospital systems with experts that lack contemporary knowledge and competence, the suggested technique is essential. A person's blood sugar level, pulse rate, age, and race are just a few medical factors that are utilised to assess whether or not they have heart disease. Dataset analyses are computed via WEKA software. [4] This study's objective was

to assess machine learning algorithms using a variety of performance metrics in order to improve accuracy. In the pre-processing stage, missing data must be replaced using the mean value. A linear kernel and SVM were employed to achieve a score accuracy of 86.8%. [5] Pathology, one of the main causes of death globally, can be reduced by having an early diagnosis. A clinical decision support system (CDSS) could be utilised to make a quick diagnosis of the participants' cardiac illness. This paper presents an effective heart disease prediction model (HDPM) for a CDSS that combines XG Boost to forecast heart disease with a hybrid Synthesized Minority Over-sampling Technique-Edited Nearest Neighbor (SMOTE-ENN) technique to equalize the spread of dataset. DBSCAN is used to detect and remove outliers.[6] The researchers have suggested the optimal machine learning approach for reliably predicting heart illness. The two feature selection techniques used are Relief and univariate extraction of features. The quality of four various types of algorithms used for machine learning has indeed been improved using cross-validation and hyperparameter optimization. The system's basic architecture can be used with Apache Kafka and Apache Spark to effectively manage Twitter data streams containing patient data. The model that fits best globally, Random Forest Classifier, has a grade of 94.9%. [7] The "Coronary endothelial Prediction Using Efficacious Machine Learning Techniques" study by Avinash Golande et al. proposes that practitioners employ a few techniques for data mining to assist law enforcement or medics in diagnosing heart disease. The most commonly used methods are Nave Bayes, k-closest, and decision trees. Other unique characterization-based methodologies often include packing computation, part width, continuous insignificant streamlining, neural nets, straight kernel self organising guidance, and SVM (Bolster Vector Machine). [8] The researchers of Machine Learning Techniques for Heart Disease, Lakshmana Rao and colleagues, expect that there will be an increase in the risk factors for heart disease (circulatory strain, diabetes, current smoker, high cholesterol, etc..). As a reason, it could be difficult to tell one type of heart disease from another. Many mining techniques and neural algorithms have been used to assess the severity of heart illness in patients. Because the definition of CHD is ambiguous, it is important to approach the issue with caution. [9]Effective Heart Disease Prognosis The authors have devised a technique to improve the accuracy of heart disease and stroke prediction by using fusion machine learning techniques. To develop the expectation model, many highlights are merged with a few in organising techniques. We construct a heart disease prediction model using a hybrid randomized forest with a linear model (HRFLM) with a standard of 88.7%. [10] For the purpose of identifying heart illness, the authors suggest a hybrid method. These algorithms' performance is assessed using the WEKA and KEEL tools. The popular dimensionality reduction method for feature extraction in the WEKA tool is PCA. Under the KEEL tool, the wrapper method is employed for feature selection. Only when the model has finished training is the chosen model tested.[11] The Heart Disease Prediction Study using Multiple Regression Model, created by K. Polaraju et al., reveals that the use of multiple linear regression is appropriate for calculating the likelihood of acquiring heart disease. The initial data set, which comprises of 3000 instances with the previously mentioned 13 different attributes, is used to complete the task. Thirty percent of the data are used for testing, while seventy percent are used for training. The outcomes demonstrate that the classification accuracy of the Regression algorithm is greater than that of other approaches. [12] In order to improve the prediction accuracy, here the authors have used hybrid approach using ensemble model. They recommended pre-processing and element selection based on an evolutionary algorithms to improve forecasting accuracy and consumption of time. The 10folds bridge technique was employed to solve the overfitting issue. Classification accuracy for the existing optimization classifier model was 98.18%. [13]Utilizing machine learning, Dr. Geetha S. and Mr. Santhana Krishnan.J. predicted heart disease. This study forecasts cardiac issues in male patients using classifying approaches. The three currently accessible WEKA interfaces—Naive Bayes, neural networks, and Decision Trees—are used in this system to forecast endothelial disease. [14] A significant portion of the global population is becoming more susceptible to heart disease as a result of lifestyle changes. It is increasingly one of the leading causes of fatalities. The healthcare professionals heavily rely on the patient's data to determine whether the patient may be at danger of heart disease because the death rate from coronary illnesses is rising.

ARCHITECTURE DIAGRAM



1. *Import packages*

Pandas is a collection of data processing and analysis tools for the Python programming language. It offers specific approaches and data structures for working with time series and mathematical tables. Thanks to the NumPy module, the Python programming language now provides a large range of advanced math operations that can be carried out on these enormous, multidimensional arrays and matrices. It is simple to generate static, animated, and dynamic visualisations using the Python programme Matplotlib. A free machine learning package called Scikit-learn is accessible for Python programmers. (formerly known as scikits.learn and also referred to as sklearn).It offers a variety of algorithms for grouping, regression, and classification, including support-vector machines, random forests, gradient boosting, k-means, and DBSCAN. Plotly is an interactive, open-source charting toolkit for Python that supports more than 40 distinct chart styles for a variety of use-cases in statistics, economics, geography, science, and tri visualisation. Use the pickle module to serialise and deserialize Python object structures. Any sort of Python object (list, dict, etc.) can be turned into byte streams by pickling, serialising, flattening, or marshalling (0s and 1s). These are the imported packages we're using for modelling.

2. *Data Preparation and Exploration*

The 14 attributes make up the dataset:

1. Age is measured in years.
2. Sex: Sex (male = 1; female = 0).
3. cp:any type of abdominal pain (Value 0: typical angina; Value 1: atypical angina; Value 2: non-anginal pain; Value 3: asymptomatic).
4. trestbps: on such being admitted to the hospital, the resting systolic in inches of mercury.
5. Cholesterol is measured in mg/dl in serum.
6. Fasting blood sugar (fbs) must be less than 120 mg/dl(1 = true; 0 = false).
7. Results of the resting electrocardiogram (rest-ECG) (Value 0: normal; Value 1: having ST-T wave abnormality; Value 2: probable or definite left ventricular hypertrophy).
8. Thalach: Thalach's heart rate peaked at this point.
9. Exercise-induced angina (exang): 1 for yes; 0 for no
10. Oldpeak:ST depression caused by exercise in contrast to idleness.
11. slope: the ST segment of the peak exercise's slope (Value 0: upsloping; Value 1: flat; Value 2: down sloping).
12. ca: the quantity of important vessels (0–3) that were stained during flouroscopy.
13. thal: chromic (normal: 3, fixed: 6, and reversible: 7).
14. The goal is endothelial disease (1 = no, 2 = yes).

All the above mentioned attributes will be generated through a blood test and used to verify the patient health status.

3. Preparing ML models

3.1. K-Nearest Neighbor:

Considering that the newest case/data and the previous cases are comparable, the K-NN method places the new example in the classification that is most similar to the other categories that are available. The K-NN method catalogues existing data using similarity and preserves all accessible information. This indicates that new data can be reliably and quickly categorised using the K-NN approach. Any regression and classification problems can be solved with the K-NN method, although classification problems are where it is most frequently used.

3.2. Random Forest Classifier:

Numerous decision trees are applied to different subsets of the provided dataset using the random forest classification algorithm, which aggregates the results to increase the dataset's projected accuracy. For ML matters surrounding both regression and classification, Random Forest may be employed. Deep learning, known as the method of merging supervised learning to solve a difficult problem and enhance the model's accuracy, is the fundamental tenet on which Random Forest is built.

3.3. Support Vector Machine:

The SVM method looks to find the optimum line or distance measure that can divide n-dimensional space into classes in order to efficiently classify fresh data points in the future. This optimal decision boundary is known as a hyper plane. The hyper plane is constructed using the extreme positions and vectors selected using SVM. Regression and classification issues are dealt with using support vector machines. The SVM algorithm may be utilized to text classification, image classification, and face detection.

3.4. Decision Tree:

The decision node and the leaf node are the two types of nodes that make up a decision tree. Decision networks are used to produce decisions and have numerous branches, as opposed to Leaf nodes, which are the outcomes of decisions and do not have any additional branches. According to the given circumstances, it offers a graphic representation of each option for resolving a problem or making a decision. The technique predicts the classes of the input dataset by starting at the root of the tree of a decision tree and working its way up. By contrasting the results of the record (actual dataset) property to those in the root property, this algorithm is applied the branch and moves on to the subsequent node.

3.5. Logistic Regression:

With this method, it is possible to predict the outcomes of a dependent variable that is categorical. The outcome must therefore be a continuous or category value. It provides stochastic values that fall around 0 and 1 rather than the exact numbers between 0 and 1. Possible results include True, False, 0 or 1, as well as Yes or No. Logistic regression is used to overcome the classification problems. A crucial machine learning technique for categorizing new data utilising both discrete and continuous datasets is logistic regression.

3.6 Naive Bayes Classifier:

This supervised training method is based on the Bayes theorem and used to address classification problems. One of the easiest and most effective ways for classifying data is the Naive Bayes Classifier, which enables the creation of machine learning models quickly and with quick prediction capabilities. Since it uses a probabilistic classifier, it based its forecasts on the possibility that a specific event will occur.

3.7 Extreme Gradient Boost:

Extreme gradient boosting is an ensemble machine learning approach that can be used to solve categorization or regression-based computational modeling issues. Extreme Gradient Boosting, often known as XGBoost, is a potent open-source variant of the gradient boosting methodology. As a result, XG Boost consists of an algorithm, an open-source project, and a Python module.

4. Models evaluation and selection

Evaluating accuracy of each model and then the best model is found. Store the model as a pickled serialised object.

5. *Web Page Development*

A Webpage containing a form is created, the form contains details to be entered by the user. The webpage uses the best model which we obtained after comparing all the models for evaluating the user entered data.

6. Prediction

The user data is utilised to forecast whether a given user detail has heart disease. It uses the best model which gives the most accuracy to evaluate for the given data. It then predicts the result.

EXPIREMENTAL RESULTS AND DISCUSSION

A. Attributes Correlations

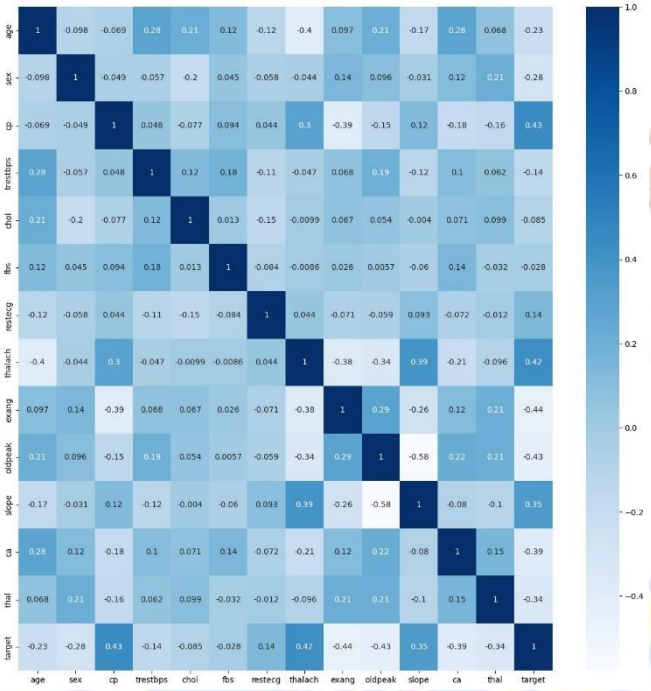


Fig 1 – Correlation chart

According to the correlation plot above, heart disease (target) is directly correlated with the type of chest pain (cp), aerobic activity angina (exang), ST depression brought on by workout in compared to rest (oldpeak), slope of the peak exercise ST segment, number of major vessels (0-3) colored by flourosopy (ca), and thalassemia (thal). Also, we notice a negative relationship between cardiac illness and maximum heart rate (thalch). Also, we see a correlation between the following characteristics: age and the quantity (0–3) of primary vessels that have been colored by flourosopy (ca). The slope of the ST section with the highest level of activity as well as the ST depression brought on by exercise in contrast to rest (oldpeak). Age, maximum heart rate, and the kind of chest pain (cp) and exercise-induced angina (exang) (thalch).

B. Logistics Regression

```
confussion matrix
[[26  4]
 [ 5 19]]

Accuracy of Logistic Regression: 83.33333333333334
```

	precision	recall	f1-score	support
1	0.84	0.87	0.85	30
2	0.83	0.79	0.81	24
accuracy			0.83	54
macro avg	0.83	0.83	0.83	54
weighted avg	0.83	0.83	0.83	54

Fig 2 – Logistic Regression accuracy

For the given dataset, Logistic Regression shows an accuracy of 83 .

C. Naive Bayes

```
confussion matrix
[[23  7]
 [ 7 17]]


Accuracy of Naive Bayes model: 74.07407407407408

              precision    recall  f1-score   support

           1       0.77      0.77      0.77        30
           2       0.71      0.71      0.71        24

    accuracy                           0.74        54
   macro avg       0.74      0.74      0.74        54
weighted avg       0.74      0.74      0.74        54
```

Fig 3 – Naïve Bayes accuracy

For the given dataset, Naive Bayes shows an accuracy of 74

D. Random Forest

```
confussion matrix
[[28  2]
 [ 6 18]]


Accuracy of Random Forest: 85.18518518518519

              precision    recall  f1-score   support

           1       0.82      0.93      0.87        30
           2       0.90      0.75      0.82        24

    accuracy                           0.85        54
   macro avg       0.86      0.84      0.85        54
weighted avg       0.86      0.85      0.85        54
```

Fig 4 – Random Forest accuracy

For the given dataset, Random Forest shows an accuracy of 85

E. *Extreme Gradient Boost*

```
confussion matrix
[[25  5]
 [ 7 17]]


Accuracy of Extreme Gradient Boost: 77.77777777777779

              precision    recall  f1-score   support

           1       0.78      0.83      0.81        30
           2       0.77      0.71      0.74        24

    accuracy                           0.78        54
   macro avg       0.78      0.77      0.77        54
weighted avg       0.78      0.78      0.78        54
```

Fig 5 – Extreme Gradient Boost accuracy

For the given dataset, Extreme Gradient Boost shows an accuracy of 77.

F. K-Neighbors Classifier

```
confussion matrix
[[23  7]
 [ 9 15]]


Accuracy of K-NeighborsClassifier: 70.37037037037037

              precision    recall  f1-score   support

           1       0.72      0.77      0.74        30
           2       0.68      0.62      0.65        24

    accuracy                           0.70        54
   macro avg       0.70      0.70      0.70        54
weighted avg       0.70      0.70      0.70        54
```

Fig 6 – K-Neighbors Classifier accuracy

For the given dataset, K-Neighbors Classifier shows an accuracy of 70

G. Decision Tree Classifier

```
confussion matrix
[[23  7]
 [ 7 17]]


Accuracy of DecisionTreeClassifier: 74.07407407407408

              precision    recall  f1-score   support

           1       0.77      0.77      0.77        30
           2       0.71      0.71      0.71        24

    accuracy                           0.74        54
   macro avg       0.74      0.74      0.74        54
weighted avg       0.74      0.74      0.74        54
```

Fig 7 – Decision Tree Classifier

For the given dataset, Decision Tree Classifier shows an accuracy of 74

H. Support Vector Classifier

```
confussion matrix
[[27  3]
 [11 13]]


Accuracy of Support Vector Classifier: 74.07407407407408

              precision    recall  f1-score   support

           1       0.71      0.90      0.79        30
           2       0.81      0.54      0.65        24

    accuracy                           0.74        54
   macro avg       0.76      0.72      0.72        54
weighted avg       0.76      0.74      0.73        54
```

Fig 8 – Support Vector Classifier

For the given dataset, Support Vector Classifier shows an accuracy of 74

I. *Accuracy Comparison of different models*

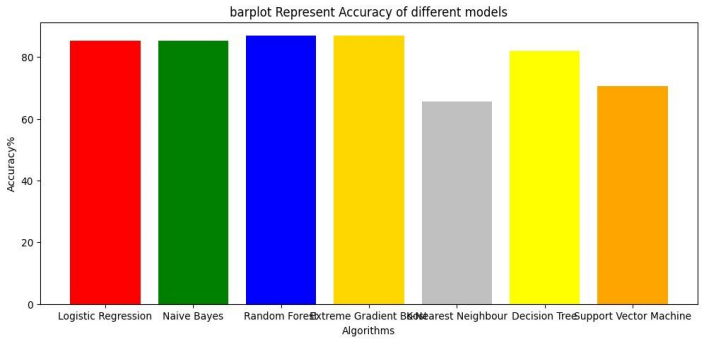


Fig 9 – Pie chart

	Model	Accuracy
0	Logistic Regression	85.245902
1	Naive Bayes	85.245902
2	Random Forest	86.885246
3	Extreme Gradient Boost	86.885246
4	K-Nearest Neighbour	65.573770
5	Decision Tree	81.967213
6	Support Vector Machine	70.491803

Fig 10 – Accuracy comparison

Comparing all models and the accuracy of Random Forest Classifier and extreme gradient booster seems to be the highest so we will select anyone as the primary classifier.

J. Webpage Outlook

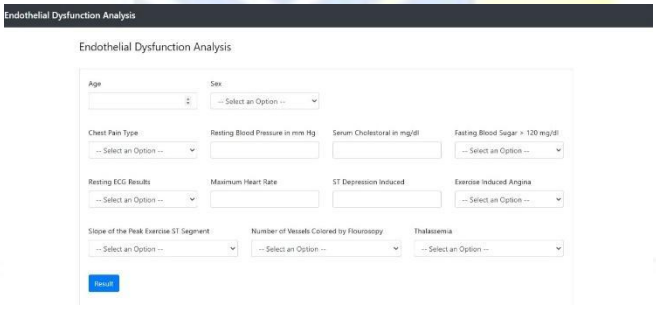


Fig – 11 Webpage

Using HTML, we have created several forms for getting inputs from users and then compared with our model and the result is displayed.

K. Prediction

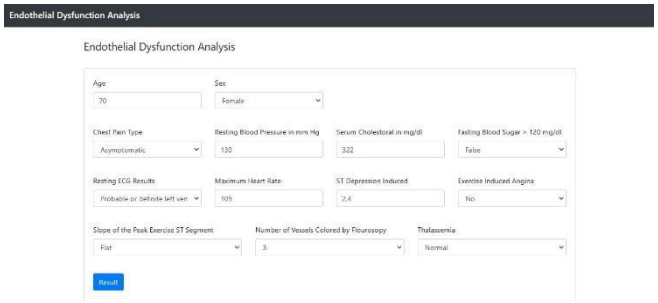


Fig – 12 Prediction (1)

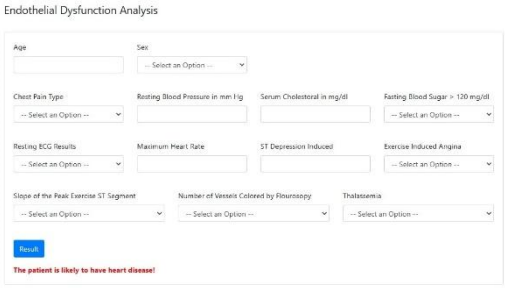


Fig – 13 Prediction (2)

From the above inputs , we have predicted that the person is having an endothelial/ heart disease.

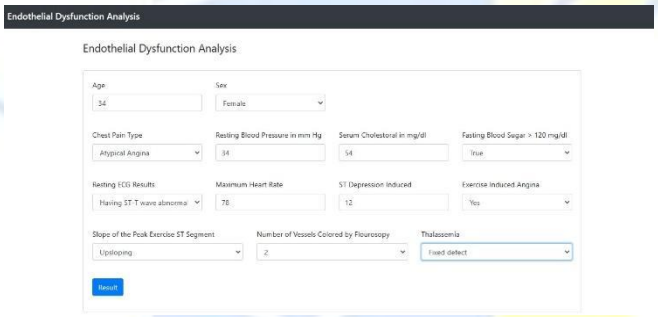


Fig 14 – Prediction (3)

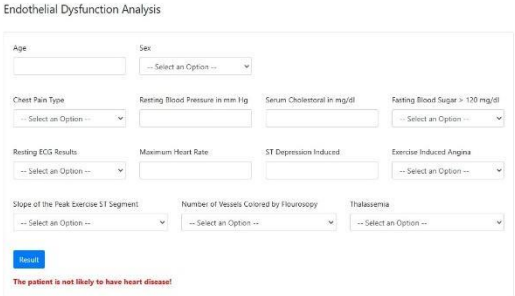


Fig 15 – Prediction (4)

From the above inputs , we have predicted that the person is not having an endothelial / heart disease.

CONCLUSION

Algorithm accuracy is one of the criteria taken into account when assessing how effectively an algorithm performs since the human heart is one of the body's most important organs and because endothelial disease prediction is a key concern for individuals. The dataset used for both testing and training purposes demonstrates the precision of machine learning techniques. The algorithm we find to be the best when we analyse algorithms based on datasets and confusion matrices is Random Forest. In the future, additional machine learning techniques will be employed to analyse endothelial dysfunction in order to forecast diseases earlier and prevent lives by raising public awareness of them.

REFERENCES

M. P. BROWN AND K. AUSTIN, APPL. PHYS. LETTERS 85, 2503–2504 (2004).

1. Abdeldjouad, Fatma Zahra et al. “A Hybrid Approach for Heart Disease Diagnosis and Prediction Using Machine Learning Techniques.” *The Impact of Digital Technologies on Public Health in Developed and Developing Countries* 12157 (2020): 299 - 306.
2. Amen, Khalid et al. “Machine Learning for Multiple Stage Heart Disease Prediction.” (2020).
3. Beyene, C. & Kamat, Pooja. (2018). Survey on prediction and analysis the occurrence of heart disease using data mining techniques. International Journal of Pure and Applied Mathematics. 118. 165-173.
4. Dengqing Zhang, Yunyi Chen, Yuxuan Chen, Shengyi Ye, Wenyu Cai, Junxue Jiang, Yechuan Xu, Gongfeng Zheng, Ming Chen, "Heart Disease Prediction Based on the Embedded Feature Selection Method and Deep Neural Network", *Journal of Healthcare Engineering*, vol. 2021, Article ID 6260022, 9 pages, 2021.
5. Fitriyani, Norma Latif et al. “HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System.” *IEEE Access* 8 (2020): 133034-133050.
6. Hager Ahmed, Eman M.G. Younis, Abdeltawab Hendawi, Abdelmgeid A. Ali, Heart disease identification from patients’ social posts, machine learning solution on Spark, Volume 111, 2020 , Pages 714-722, ISSN 0167-739X.
7. International Journal of Recent Technology and Engineering (IJRTE)ISSN: 2277-3878, Volume-8, Issue-1S4, June 2004.
8. Lakshmanarao, Aravind, Y.Swathi and P.Sri Sai Sundareswar. “Machine Learning Techniques For Heart Disease Prediction*.” International Journal of Scientific & Technology Research* 8 (2019): 374-377.
9. Mohan, Senthilkumar, Chandrasegar Thirumalai and Gautam Srivastava. “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques.” *IEEE Access* 7 (2019): 81542-81554.
10. Nagavelli, Umarani et al. “Machine Learning Technology-Based Heart Disease Detection Models.” *Journal of healthcare engineering* vol. 2022 7351061. 27 Feb. 2022, doi:10.1155/2022/7351061
11. Polaraju, K. and Deepak Prasad. “Prediction of Heart Disease using Multiple Linear Regression Model.” (2017).
12. S. E. A. Ashri, M. M. El-Gayar and E. M. El-Daydamony, "HDPF: Heart Disease Prediction Framework Based on Hybrid Classifiers and Genetic Algorithm," in IEEE Access, vol. 9, pp. 146797-146809, 2021:10.1109/ACCESS.2021.3122789.
13. S. K. J. and G. S., "Prediction of Heart Disease Using Machine Learning Algorithms," 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), Chennai, India, 2019, pp. 1-5, doi: 10.1109/ICIICT1.2019.8741465.
14. Swain, Debabrata et al. “An Efficient Heart Disease Prediction System Using Machine Learning.” (2020).