2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)
Amity University, Noida, India. Mar 14-15, 2024

2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) | 979-8-3503-5035-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICRITO61523.2024.10522122

Supervised Machine Learning in Cardiology: A Predictive Model for Heart Disease

Anamika Mall
Department of Computer Science and Engineering
KIET Group of Institutions
Ghaziabad, UP, India
anamikamall221@gmail.com

Ankita Singh
Department of Computer Science and Engineering
KIET Group of Institutions
Ghaziabad, UP, India
ankitasingh0030122001@gmail.com

Ayush Bansal
Department of Computer Science and Engineering
KIET Group of Institutions
Ghaziabad, UP, India
bansalabayush2024@gmail.com

Hriday Kumar Gupta
Department of Computer Science and Engineering
KIET Group of Institutions
Ghaziabad, UP, India
hridaykumargupta@gmail.com
ORCID- 0000-0001-6115-225X

***Abstract*—One of the worst diseases in the world, heart disease takes a great toll on lives every year. In order to treat this potentially fatal illness, a tool or gadget must be practical, accurate, and dependable. This will allow for prompt detection or prediction, which will lead to efficient treatment and a decline in death rates. The heart is the most important organ in the human body, second only to the brain, and it is vital to the circulation of blood to all organs.A swift and accurate diagnostic method is imperative to mitigate the high mortality associated with cardiac disorders. Predicting the incidence of these disorders holds paramount importance in the medical domain. In the contemporary era, machine learning techniques prove invaluable for predicting and automating the interpretation of extensive and intricate datasets in various fields, including medicine.This research introduces a statistical model for heart disease, aiding medical examiners and cardiac practitioners in forecasting based on fundamental aspects of a patient's health history. The model, constructed using a Decision Tree, achieves an impressive accuracy of approximately 97%. The crucial functions of machine learning, data analytics, and data mining highlight how important these technologies are to improving the diagnosis and treatment of cardiac disorders.**

Keywords—Machine Learning, Logistic Regression, Decision Tree, Random Forest Classifier, Support Vector Machine

I. INTRODUCTION

Since it delivers blood to every part of the body, the heart is vital. Failure will shut down the brain and other organs, killing the person in minutes. From 1990 to 2013, cardiovascular disease mortality increased 41%, from 12.3 million to 17.3 million. Furthermore, the same problem accounts for half of all deaths in the US and other industrialized nations[1]. Lifestyle changes, work stress, and poor diets are increasing heart disease rates. The human mind cannot anticipate illness and handle massive amounts of information at once. Data mining has shown to be quite good at predicting a broad range of outcomes for multiple fields. Numerous models have been developed to predict specific events so that data mining and deep learning can be applied to them. ***The Dataset for Heart Disease research was used to train a model with four different machine-learning classifier algorithms to produce the most accurate heart disease prognosis.***

In the medical field, machine-learning is essential. To predict sickness, it uses large databases and past medical records. These days, machine learning algorithms and techniques are used to identify cardiac disease as per risk factors and the clinical history of the patient. Parameters include heart rate, age, blood pressure (BP), sex, obesity, and other factors. To compare the characteristics and forecast cardiac disease, algorithms such as Decision trees, SVMs, logistic regression, and random forest classifiers are used.

The data is divided into two sections to train and test the models. The heatmap makes it clear that factors like age, cholesterol, and old peaks are among the primary causes of heart disease.

II. LITERATURE REVIEW

The literature on decision tree-based heart disease prediction indicates an increasing awareness of the need for precise and effective models to tackle this important global health concern. Heart disease causes worldwide morbidity and death, thus scientists are using machine learning techniques more and more to improve prediction. Decision trees have become well-known in this field due to their interpretability and simplicity. Research has demonstrated the value of early detection in mitigating the consequences of heart disease, as conventional risk assessment models are ill-suited to address the complexities of cardiovascular health. Decision Trees, through their tree-like structures and recursive partitioning of data, prove effective in identifying relevant patterns and relationships within complex datasets. Researchers have conducted comparative analyses, assessing the performance of Decision Trees against other machine learning algorithms, emphasizing factors like sensitivity, specificity, and computational efficiency. Additionally, the ability of Decision Trees to perform feature selection during model-building enhances interpretability by pinpointing influential variables. Challenges, such as overfitting and handling imbalanced datasets, are acknowledged in the literature, underscoring the ongoing efforts to refine and improve the application of Decision Trees for heart disease prediction. As research progresses, addressing these challenges and exploring innovative approaches will likely contribute to the broader adoption of Decision Trees in cardiovascular health prediction, providing valuable tools for healthcare professionals.

The diagnosis of heart problems is a current field of study. There are some different techniques used to strengthen the process's efficacy and accuracy on a wide range of factors. Using the clinical dataset's decision tree, research on the heart disease prediction model was conducted in 2013. [1].

Using the CHDD dataset, more investigation on the estimation of heart disease using a combination of SVM, DT, and logistic regression was carried out that year[2].DT's rule-based algorithm benefited from the strategy, which is based on categorization, regression, and correlation. The catch is that SDL has no way of guaranteeing superior outcomes, and the outcomes are determined. An innovative machine-learning heart disease prediction technique [3] was applied on rank in a recent study conducted in 2018. Because of technological advancements, ML is now a developing discipline.The MLP algorithm is used to increase productivity and accuracy, however, it might get stuck at extremely low or high levels.

Data mining techniques were suggested for the prediction and analysis of heart disease occurrence by Chala Beyene et al [1]. The main goal is to forecast the start of cardiac problems so that they can be automatically identified early and efficiently treated. A healthcare organisation staffed by professionals who have lost their expertise would likewise benefit greatly from the suggested technique. Numerous medical factors, including blood pressure, blood sugar, heart rate, age, and sex, are used to assess whether or not cardiac illness is present.

To predict cardiac disease, hybrid machine learning was used by Senthilkumar Mohan et al [4]. The Cleveland data set is the one that was utilized. Raw data processing is the initial step. Tuples are removed in this way by using the data set that is lacking the values. Additionally, because the authors believe that the age and sex attributes in the dataset are sensitive and unrelated to the prediction method, they decide not to use them. Important clinical data is stored in the remaining 11 characteristics.

III. OBJECTIVE

The present study aims to create and test a viable cardiac disease detection model using the Decision Tree method. The primary focus is on developing an accurate and interpretable predictive tool that enables early identification of individuals at risk of heart disease. To achieve this, the project will begin with the comprehensive acquisition and preprocessing of health datasets, ensuring that diverse patient demographics, clinical parameters, and lifestyle factors are adequately represented. Subsequently, feature selection and engineering will be conducted to enhance the model's discriminatory power by identifying key variables influencing cardiac health. The implementation and training of the Decision Tree model will follow, with meticulous hyperparameter tuning to optimize accuracy, sensitivity, and specificity. Emphasis will be placed on ensuring the interpretability of the model, providing transparent insights into the factors guiding its predictions. Rigorous validation procedures, including cross-validation and external validation, will assess the generalizability of the Decision Tree model. Collaborative efforts with healthcare practitioners will facilitate the integration of the model into clinical workflows, and user feedback will be actively solicited to refine usability and address practical considerations. Ultimately, this research aims to deliver a tailored and effective Decision Tree-based solution for heart disease detection, contributing to improved patient outcomes and advancing cardiovascular health diagnostics in real-world healthcare settings.

Easy to use: A user-friendly platform is the major goal of the present study. By providing patient medical facts, the algorithm will detect heart disease on extracted attributes. A well-trained model is less likely to make mistakes when predicting the type of heart disease, which improves accuracy and saves time. It also helps doctors and patients assess heart disease risk, which can be difficult to perform on one's own.

Without human involvement: Detecting heart disease requires medical details like age and cholesterol, and the algorithm provides results based on extracted features, minimizing errors and saving time for patients and doctors. This applies when results are delivered faster. This can speed up the heart treatment process by saving doctors and patients time, allowing them to focus on other therapies and preventative measures to minimize the impact of the condition.

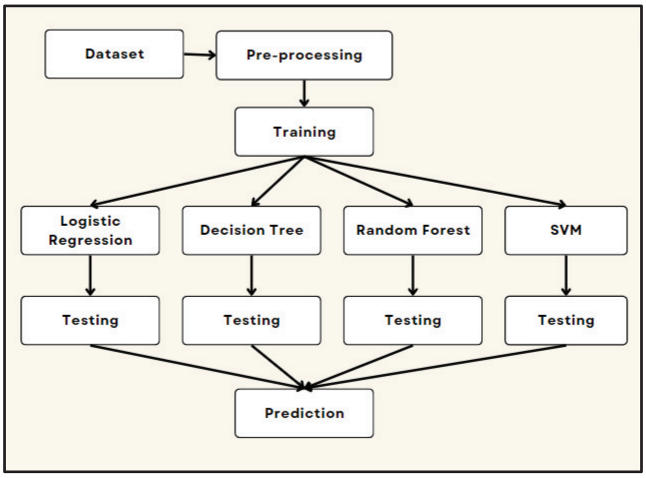


Fig. 1. Flow Chart

IV. DATASET

In this section,the dataset the authorsused to construct a machine learning-based cardiovascular prediction model is explained in detail. The dataset contains 303 samples and contains 14 input features important for cardiovascular prediction. These factors encompass a wide range of demographic, clinical, and physiological characteristics, each of which contributes to the predictive accuracy of the system. 13 clinical variables of the 303 patients in this dataset assist us in assessing the patients' potential risk for cardiovascular disease. Of the 303 samples, 165 cases represent individuals with heart failure, and the remainder (138 items) represent individuals with normal heart failure.

Below detailed information on each aspect is provided:

1. Age (Age): This element indicates the individual's age, stated in years. Given that the likelihood of heart-related issues tends to rise with age, age plays a significant role in the diagnosis of heart disease.
2. Sex (male and female): The individual's gender is entered as a binary variable, with 0 representing

female and 1 representing male. Sex may be an important predictor of cardiovascular disease risk.

3. Chest pain characteristics (cp): Chest pain is classified into four types (0, 1, 2, 3), each reflecting a different degree of severity. It is a major symptom associated with heart disease. (0: typical angina, 1: atypical angina, 2: non-anginal pain, 3: asymptomatic)

4. Resting blood pressure (trestbps): A person's resting blood pressure, expressed in mm Hg, is represented by this unit. Heart disease may be indicated by elevated blood pressure.

5. Serum cholesterol (chol): A reliable indicator for cardiovascular disease, serum cholesterol levels are expressed in mg/dl. Heart disease can result from high cholesterol levels.

6. Fasting blood sugar (fbs): A person's blood sugar level above 120 mg/dl can be determined using this binary (0 or 1) signal. Elevated fasting blood sugar levels are generally associated with cardiovascular risk.(> 120 mg/dl; true = 1; false = 0)

7. Resting electrocardiogram results (restecg): This feature allows the recording of resting electrocardiogram findings. It takes three values (0, 1, 2), reflecting various abnormalities that can indicate cardiac problems.(0: normal, 1: abnormal ST-T waves, and 2: likely or definitely exhibiting left ventricular hypertrophy according to Estes' criterion)

8. Maximum heart rate attained (thalach): The term "thalach" refers to the highest heart rate that can be reached when exercising. This is a crucial physiological measure related to heart failure.

9. Exercise angina (exang): This binary variable (0 or 1) indicates whether the person suffers from exercise-induced angina. Angina can be a symptom of coronary artery disease. (1: yes, 0: no)

10. Exercise-induced ST depression relative to rest (oldpeak): Oldpeak represents the degree of exercise-induced ST-side depression in relation to rest. It is employed to gauge the heart's stress during physical activity.

11. Slope of peak exercise ST Segment (slope): The three values (0, 1, 2) that correspond to the form of the ST segment and are appropriate for cardiac diagnostic applications are derived from the slope of the peak workout ST segment. (0: upsloping, 1: flat, 2: downsloping)

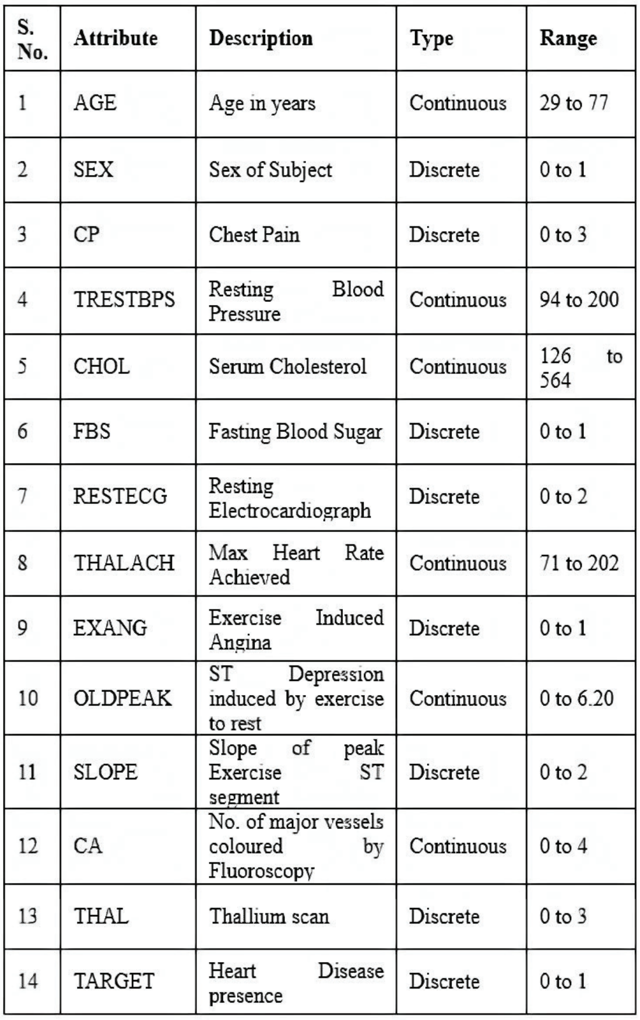
12. Number of large vessels stained by fluoroscopy (ca): This is a categorical variable that shows how much coronary disease has been affected by the number of large blood vessels stained by fluoroscopy (0–3)

13. Thalassemia (thal): Thalassemia is another variant in groups representing different forms of thalassemia. It can be linked to heart disease. (3: normal; 6: fixed defect; 7:reversible defect)

14. Target(s): The target is a binary variable (0 or 1) that indicates if heart failure is present (1) or not (0). This is the outcome variable that a machine learning model aims to predict.

The pharmaceutical sector provides these datasets, which contain 303 samples and 14 inputs. Many machine learning methods for cardiovascular disease prediction are trained and evaluated using these datasets. Thorough evaluation of cardiovascular risk can be achieved, and predictive models are rigorously evaluated. The data is split into two sections to train and test the models. The training and testing sections make up theserecords.

TABLE I. DATASET AND RANGE



S. No.	Attribute	Description	Type	Range
1	AGE	Age in years	Continuous	29 to 77
2	SEX	Sex of Subject	Discrete	0 to 1
3	CP	Chest Pain	Discrete	0 to 3
4	TRESTBPS	Resting Blood Pressure	Continuous	94 to 200
5	CHOL	Serum Cholesterol	Continuous	126 to 564
6	FBS	Fasting Blood Sugar	Discrete	0 to 1
7	RESTECG	Resting Electrocardiograph	Discrete	0 to 2
8	THALACH	Max Heart Rate Achieved	Continuous	71 to 202
9	EXANG	Exercise Induced Angina	Discrete	0 to 1
10	OLDPEAK	ST Depression induced by exercise to rest	Continuous	0 to 6.20
11	SLOPE	Slope of peak Exercise ST segment	Discrete	0 to 2
12	CA	No. of major vessels coloured by Fluoroscopy	Continuous	0 to 4
13	THAL	Thallium scan	Discrete	0 to 3
14	TARGET	Heart Disease presence	Discrete	0 to 1

A. Histogram of Attributes

The dataset's range of properties is displayed by the attributes histogram [10].Using a histogram, which bins the data, is the fastest approach to see the distribution of each property. The code which is used to create it is:

dataset.hist()

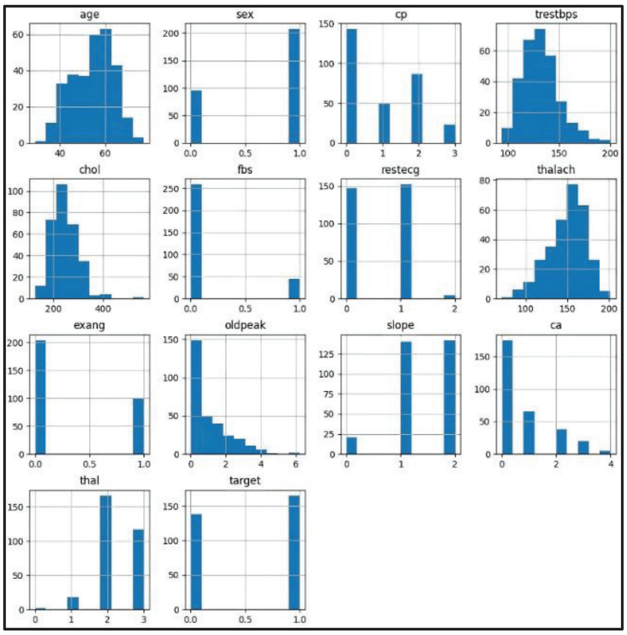


Fig. 2. Histogram

V. IMPLEMENTED WORK AND ALGORITHM

The authors have looked at four different machine learning methods in our research study: support vector machine (SVM), random forest classifier, decision tree, and logistic regression. The authors refrained from proceeding with a novel approach in experimentation due to the inherent need for validating the necessity of employing multiple machine learning algorithms. The decision to utilize well-established algorithms such as Decision Tree, Logistic Regression, SVM, and Random Forest was deliberate, aimed at benchmarking their performance and evaluating their efficacy in the context of cardiac disease detection. This approach ensures a robust comparative analysis and contributes to the broader understanding of the strengths and limitations of these established algorithms in the specific domain under investigation.

A. Logistic Regression

The aim of the supervised machine learning technique known as logistic regression is to estimate the likelihood that an instance will belong to a particular class. It is primarily used in classification issues. A popular classification method used in many different fields is logistic regression. This procedure is called "regression" because it takes the output of the linear regression function and uses a sigmoid function to calculate the likelihood that the given class carries.The corresponding outputs of logistic regression and linear regression are a key point of differentiation. While logistic regression is intended to determine the likelihood that an instance will belong to a particular class or not, linear regression generates a continuous result that can span a large range of numerical values.

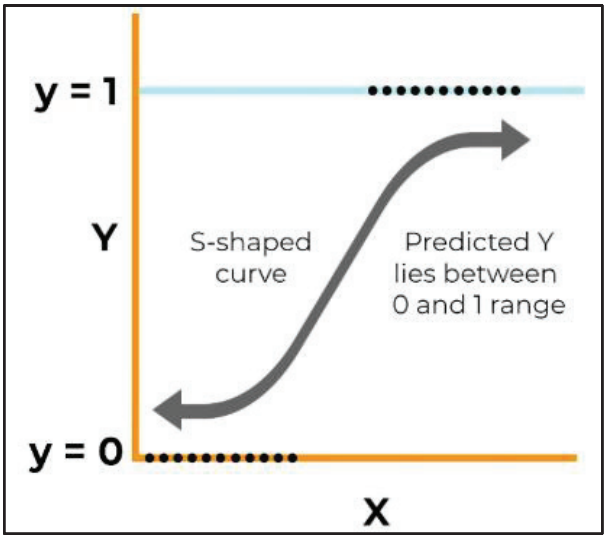


Fig. 3. Logistic Regression

B. Decision Tree

Conversely, decision trees are supervised machine learning methods, that represents the data graphically.[10]It is a non-parametric approach for supervised learning in regression and classification.Its hierarchical tree structure has internal, leaf, branch, and root nodes.In machine learning, a decision tree is a way to make decisions by asking a series of questions about your data. Each question is based on a specific feature (a characteristic of your data), and the answer to each question leads you to a decision. These questions and answers form a tree-like structure.Internal nodes represent attribute tests, branches indicate outcomes, and leaves represent decisions made after processing in a decision tree. [7]

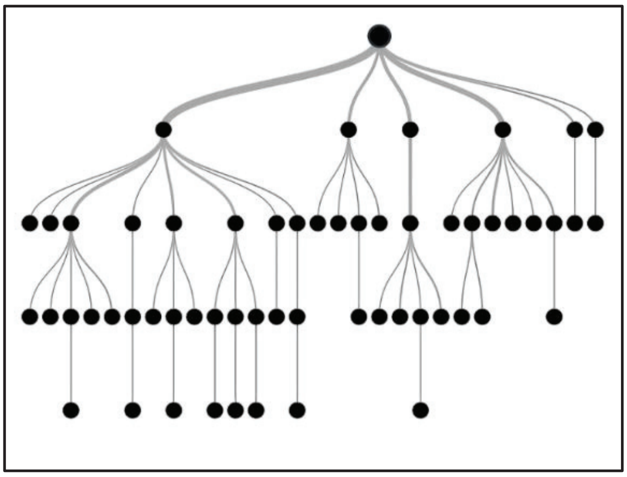


Fig. 4. Decision Tree

C. Random Forest Classifier

Another supervised machine learning algorithm is Random Forest. Although this method works better in classification problems, it can be utilized in regression tasks as well. [9]A single choice is produced by combining several decisions using Random Forest regression. [7]Random Forest uses numerous decision trees together. The class with the highest votes is predicted by our model. The decision tree algorithm's limitations are eliminated since every random forest tree forecasts a class. By doing this, accuracy is increased and the dataset's overfitting is decreased. [11]We make sample datasets for each model by picking rows and traits at random from the dataset. This part is called Bootstrap.

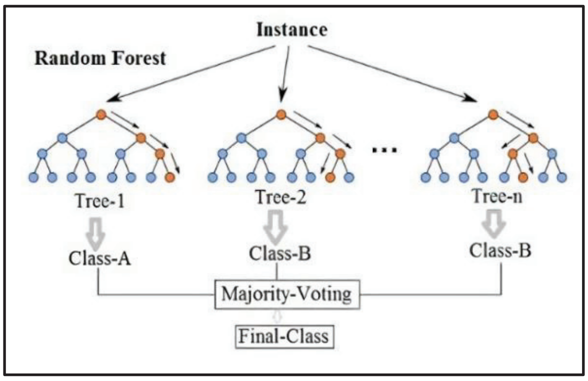


Fig. 5. Random Forest

D. *Support Vector Machine (SVM)*

Using a pre-set goal variable, the Support Vector Machine is a famous guided machine learning method that can be applied as both a predictor and a classifier. [9]In this approach, the data is set by an optimal hyperplane that separates all the points of data from one class from those of the other class. The model is thought to be better when the two groups have a bigger gap or edge. [8]As is customary in this setting, SVM is applied to two distinct datasets: a training dataset and a test dataset. The classes are linearly separable in an ideal world. There is a line that precisely split into two groups in such a case. Nevertheless, a number of lines combine to split the dataset precisely rather than just one. The best line is chosen to serve as the "separating line" among these. [5]

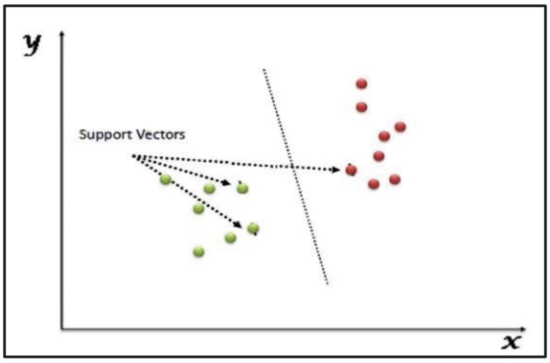


Fig. 6. Support Vector Machine

VI. RESULT ANALYSIS

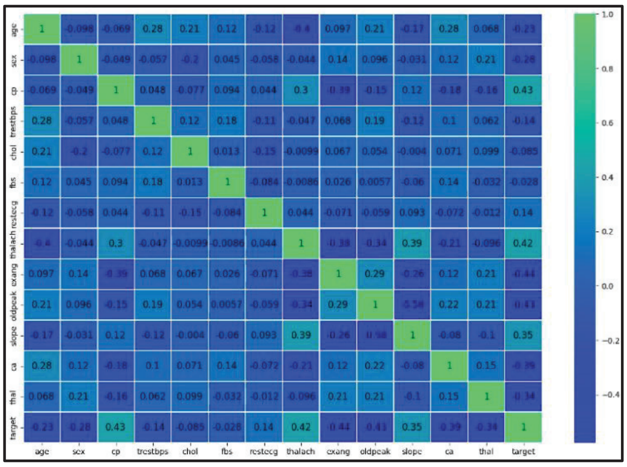


Fig. 7. Heatmap

The heatmap clearly illustrates that age, cholesterol, and old peaksthese things make you more likely to get heart disease.

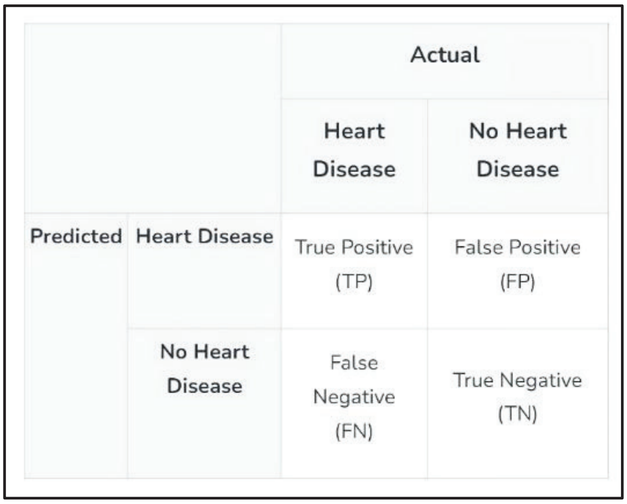
As you can see from the image above, there is a link between the target and chest pain (cp). It means that heart disease is more likely to strike people who have a high risk of experiencing chest pain. Thalachal, slope, and resting have good correlations with the target in addition to chest discomfort.

After that, there is a bad link among the aim and exercise-induced angina (exang), meaning that while we exercise, our hearts need more blood, but our narrower arteries reduce that blood supply. Old peak and thal, in addition to ca, exhibit a negative connection with the target.

Accuracy Calculation

Algorithm accuracy is determined by four values: TP, FP, TN, and FN. These numbers are obtained from the confusion matrix.

A confusion matrix gives you an idea of how well a model developed using machine learning did on a collection of test data. It is frequently employed check to see how well classification models perform, with the goal of predicting the presence of a category label in each input instance. The number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) that the model produced using the test data is displayed in the matrix.



		Actual	
		Heart Disease	No Heart Disease
Predicted	Heart Disease	True Positive (TP)	False Positive (FP)
	No Heart Disease	False Negative (FN)	True Negative (TN)

Fig. 8. Confusion matrix

The following is another method that a confusion matrix can be displayed as a matrix: [6]

[[TP FP

FN TN]]

The following formula can be used to determine the algorithm's accuracy:

Accuracy = {(TP + TN) / TP + FP + TN + FN)} * 100

A. *Logistic Regression*

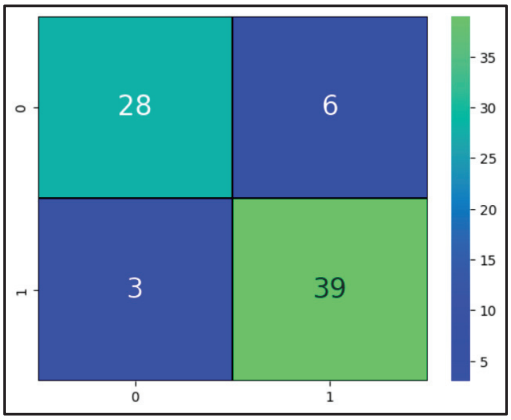


Fig. 9. Confusion Matrix for Logistic Regression

Conclusion: Based on the previous data, the authors can infer that the Logistic Regression classifier has an accuracy of approximately 88%.

B. *Decision Tree*

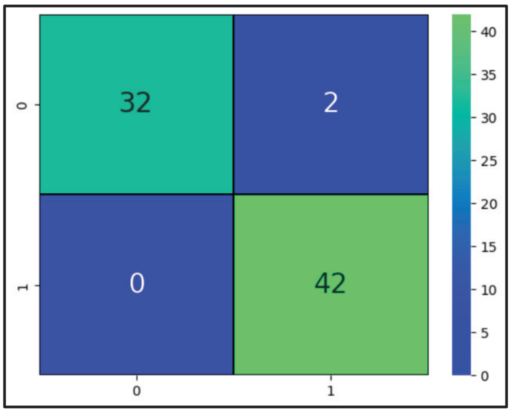


Fig. 10. Confusion Matrix for Decision Tree

Conclusion: Based on the previous data, the authors can infer that the Decision Tree classifier has an accuracy of approximately 97%.

C. *Random Forest Classifier*

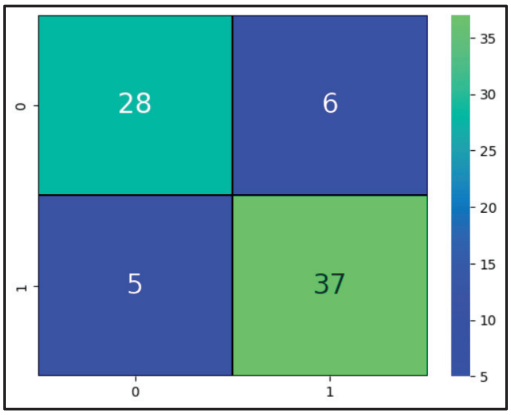


Fig. 11. Confusion Matrix for Random Forest

Conclusion: Based on the previous data, the authors can infer that the Random Forest classifier has an accuracy of approximately 85%.

D. *Support Vector Machine*

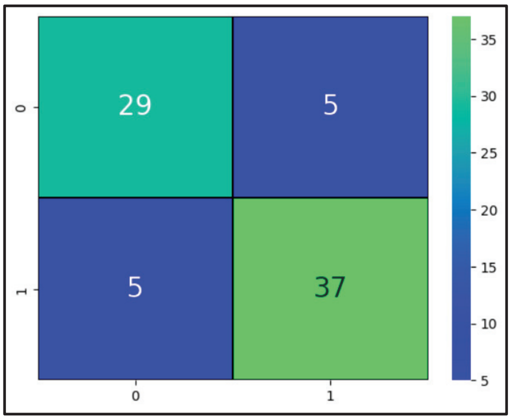


Fig. 12. Confusion Matrix for Support Vector Machine

Conclusion: Based on the previous data, the authors can infer that the Support Vector Machine classifier has an accuracy of approximately 86%.

After testing and training with machine learning, the decision tree beats other methods in accuracy. Calculate each method's accuracy using the confusion matrix.

TABLE II. ACCURACY TABLE

Algorithm	Accuracy	Sensitivity	Specificity	Precision
Logistic Regression	88%	90%	86%	82%
Decision Tree	97%	100%	95%	94%
Random Forest Classifier	85%	84%	86%	82%
Support Vector Machine	86%	85%	88%	85%

VII. CONCLUSION

The accuracy of an algorithm is determined since the heart is one of the body's critical organs and heart disease prediction is a significant concern. The dataset that is utilized for training and testing determines how accurate the machine learning algorithms are[10]. The most efficient approach is chosen by Decision Tree analysis of the dataset and confusion matrix. With over 97% accuracy, the Decision Tree strategy is best for our model.

Now that the authors have the best working algorithm (Decision Tree Classifier) applied to our model, the authors can use the available data to determine whether or not our model will produce the correct output.

REFERENCES

[1] Beyene, Chala, and Pooja Kamat. "Survey on prediction and analysis of the occurrence of heart disease using data mining techniques." *International Journal of Pure and Applied Mathematics* 118.8 (2018): 165-174.

[2] Mythili, T., et al. "A heart disease prediction model using SVM-decision trees-logistic regression (SDL)." *International Journal of Computer Applications* 68.16 (2013).

[3] Robertson, Cassandra Burke, and Sharona Hoffman. "Professional speech at scale." *UC Davis L. Rev.* 55 (2021): 2063.

[4] Mohan, Senthilkumar, Chandra segar Thirumalai, and Gautam Srivastava. "Effective heart disease prediction using hybrid machine learning techniques." *IEEE Access* 7 (2019): 81542-81554.

[5] Rindhe B. U., Ahire N., Patil R., Gagare S., &Darade M. (2021). Heart Disease Prediction Using Machine Learning. International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), 5(1), 267-273. DOI: 10.48175/IJARSCT-1131

[6] Garg A., Sharma B., & Khan R. (2021). Heart disease prediction using machine learning techniques. IOP Conf. Series: Materials Science and Engineering, 1022(1), 012046. doi:10.1088/1757-899X/1022/1/012046

[7] Kavitha M., Gnaneswar G., Dinesh R., Sai Y. R., & Sai Suraj R. (2021). Heart Disease Prediction Using Hybrid Machine Learning Model. In Proceedings of the Sixth International Conference on Inventive Computation Technologies (ICICT 2021) (pp. 1329-1333). IEEE Xplore. Part Number: CFP21F70-ART. ISBN: 978-1-7281-8501-9. DOI: 10.1109/ICICT50816.2021.9358597

[8] Sharma V., Yadav S., & Gupta M. (2020). Heart Disease Prediction using Machine Learning Techniques. In Proceedings of the 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) (pp. 177-181). IEEE. DOI: 10.1109/ICACCCN51052.2020.9362842

[9] Ramalingam V. V., Dandapath A., & Karthik Raja M. (2018). Heart disease prediction using machine learning techniques: a survey. International Journal of Engineering & Technology, 7(2.8), 684-687. DOI: 10.14419/ijet.v7i2.8.10557.

[10] Singh A., & Kumar R. (2020). Heart Disease Prediction Using Machine Learning Algorithms. In Proceedings of the 2020 International Conference on Electrical and Electronics Engineering (ICE3-2020) (pp. 452-457). IEEE. DOI: 10.1109/ICE348803.2020.9122958

[11] Bhatt C. M., Patel P., Ghetia T., & Mazzeo P. L. (2023). Effective Heart Disease Prediction Using Machine Learning Techniques. Algorithms, 16, 88. https://doi.org/10.3390/a16020088

[12] Lakshmanarao, A., Swathi, Y., &Sundareswar, P. S. S. (2019). Machine learning techniques for heart disease prediction. *Forest*, *95*(99), 97.

[13] Srivastava, K., & Choubey, D. K. (2020). Heart disease prediction using machine learning and data mining. *International Journal of Recent Technology and Engineering*, *9*(1), 212-219.