Cardiovascular Disease Prediction based on Decision Tree

R Bhuvaneswari
Department of Computer Science and Engineering
Amrita School of Computing, Amrita Vishwa Vidyapeetham
Chennai, India
bhuvanacheran@gmail.com

S Karthigeyan
Department of Computer Science and Engineering
Amrita School of Computing, Amrita Vishwa Vidyapeetham
Chennai, India
karthigsk693@gmail.com

***Abstract*—This study uses advanced data analytics approaches to explore the vital field of heart disease, a prevalent and potentially fatal disorder. The study uses supervised machine learning to identify predicted patterns in people's vulnerability to heart disease, primarily concentrating on the Decision Tree (DT) method. The study is taking place in the context of rising mortality rates from cardiovascular diseases, which highlights the critical need for preventative healthcare measures. Notably, the DT algorithm proves to be an exceptional performer, predicting the risk of heart disease with an astounding 99% accuracy. This supports its usefulness as a diagnostic tool and highlights how revolutionary preventative healthcare may be with its potential. The work adds to the growing conversation on the use of technology to solve pressing healthcare issues in addition to illuminating the predictive power of machine learning in cardiovascular health. The research's conclusions have ramifications for data-driven healthcare treatments and personalized therapy in the future, opening the door to creative ways to treat cardiac disease.**

***Index Terms*—Decision Tree, Cardiovascular Disease, Supervised learning, Diabetes, Machine learning model.**

I. INTRODUCTION

Heart disease is difficult to diagnose because of various health problems including Diabetes, high blood pressure, excess cholesterol and irregular heart rate. Multiple Data interpretation strategies were used and the approach taken determines the severity. of heart disease in individuals. Disease severity is classified using several techniques, including K-NearestNeighbor (KNN) algorithm, DT and NaiveBayes (NB) algorithm. Because of the complexity of heart disease, it must be treated with caution. Failing in this case can affect our heart or lead to early death. Medical and statistical perspectives are used to identify different types of metabolic disorders. Data mining is the process of extracting needed information from huge databases in various fields encompassing business, healthcare, and education. One includes machine learning (ML) area in artificial intelligence (AI) which is developing at a rapid pace. These algorithms are able to analyze numerous amounts of data from many different fields, one of which is medical fields. It replaces the conventional predictive modeling approach, which uses a host for gaining knowledge of the complex and non-linear interactions of a wide range of factors, reducing the discrepancy between the predicted and actual outcomes. It is called data mining to sort through vast data sets to obtain significant decision-making data from a gathering of primary source documents for upcoming studies. The doctor is full of patient information. This information must be assessed using various machine learning techniques. Also Healthcare professionals analyze this information. Through analysis, medical data mining with classification algorithms provides therapeutic assistance. It evaluates the methods used to classify patients at risk of heart disease. In the UCI laboratory, data from heart patients are used for pattern recognition using DT, Support Vector Machines (SVM), and Naive Bayes. Various algorithms accuracy and performance are compared. The hybrid approach that is suggested obtains an F-measure accuracy of 99%, and it can be compared to other existing methods. This study shows that DT attains the highest accuracy and that they can be improved in terms of efficiency by different methods and parameter tuning. This study compared classification algorithms DT, LR, NB and SVM in terms of accuracy. The decision tree algorithm was the most accurate. Naive Bayes, KNN, DT were used as classification methods and the classifiers accuracy was assessed in the scale of attributes.

This research turns out to be unique as an accuracy of 99% is obtained which is higher than the previous ones.

II. RELATED WORK

Because cardiovascular diseases (CVDs) are a major danger to global health, there is an urgent need for improved early detection and prognosis techniques. Revolutionary advancements have been made in the application of machine learning (ML) to cardiovascular research, providing hitherto unseen insights and prediction power. Examining important research publications from 2023, this review sheds light on the state-of-the-art developments in ML-driven cardiovascular illness prediction. ML applications for cardiovascular disease prediction are thoroughly examined by Subramani S et al. (2023). Their work presents a sophisticated investigation of the topic by illuminating the combination of machine learning (ML) and the Internet of Things (IoT) to detect cardiovascular disease (CVD) using cardiac sounds [1]. Age-related modifications are shown when deep learning and tree-based models are applied to analyze electrocardiogram (ECG) data. Explainable AI is used to identify discriminative characteristics by analyzing

ECG features and raw signals in a heterogeneous dataset. The results show reductions in certain ECG characteristics indicative of age and inferred breathing rates, providing fresh perspectives beyond conventional methods [2].Following cardiovascular illnesses in terms of worldwide mortality, cancer is a major global health concern. Over 7.8 million women worldwide will be affected by breast cancer by 2020, according to the World Health Organization, making it the most common malignancy. In order to forecast the stages of breast cancer, Bonaventure F. P. Dossou et al.(2023) made a study that uses computer vision models that have already been trained using the Nightingale Open Science dataset of breast biopsy photos. Although individual models perform well, our results demonstrate the higher effectiveness of an ensemble model, providing a workable solution [3]. An in depth review of machine learning techniques for cardiovascular disease prediction using big data from the medical field is provided by Baghdadi NA et al. (2023). This work contributes to the ongoing attempts to improve prediction capacities by exploring advanced machine learning approaches [4]. In 2022, Bruno Machado Pacheco conducted a meta-analysis examining the use of machine learning in cardiovascular disease prediction, providing valuable insights. The study delves into the prediction of brain age using neuroimaging data and emphasizes the significance of pre-training deep learning models for accurate brain age estimation. Pacheco's novel approach involves pre-training models on tasks related to the brain, resulting in cutting-edge outcomes. The research validates the versatility of deep learning models across various health conditions [5]. A comprehensive meta-analysis of machine learning applications for cardiovascular disease prediction is presented by Chinnasamy P et al. in 2022. Despite being published in 2022, it offers insightful information about the state of machine learning in cardiovascular research overall [6]. Youssef Fakir et al.(2022) delve into the efficacy of Data Mining (DM) techniques for information extraction from extensive datasets. The research scrutinizes various classification methods, including Decision Tree (DT), C-RT, C5.0, AD-Tree, and CS-MC4 algorithms. Evaluation metrics such as Recall, precision, and F-measure are employed to assess their performance. Notably, the study highlights the effectiveness of the AD-Tree algorithm, demonstrating superior speed and accuracy when applied to a Diabetes dataset. This research contributes valuable insights into the diverse applications of DM algorithms, emphasizing the importance of choosing appropriate techniques for classification tasks. [7]. Mohamed G. El-Shafiey et al.(2022) tackle the worldwide issue of cardiac ailments and provide a novel method for early detection. The paper presents GAPSO-RF, an optimized technique based on random forest (RF) that combines the advantages of hybrid genetic algorithms (GA) and particle swarm optimization (PSO). To improve the accuracy of heart disease prediction, GAPSO-RF concentrates on feature selection. The method combines local search with PSO, global search with modified GA, and discriminate mutation strategy in GA. It also makes use of multivariate statistical analysis. GAPSO-RF outperforms other cutting-edge techniques when tested using datasets from the University of California (Cleveland and Statlog), with high prediction accuracies of 95.6% and 91.4%, respectively. The development of algorithms for predicting heart disease is greatly aided by this research [8]. T. Poojitha and R. Mahaveerakannan(2023) compare the Estimation Precision in Random Forest and Logistic Regression for coronary disease. Utilizing 143 samples, the study reveals that the novel Random Forest outperforms Logistic Regression, achieving an accuracy of 90.16% compared to 85.25%. The statistical test between the two classifiers is significant (p ¡ 0.05). This study highlights the superior predictive capabilities of the Random Forest algorithm for coronary disease [9]. Using electronic health data, Boukhatem et al.(2022) investigate machine learning techniques for precise cardiac disease prediction. Using main health characteristics, the study uses four classification methods: Random Forest (RF), Support Vector Machine (SVM), Multilayer Perceptron (MLP), and Naïve Bayes (NB). The data is preprocessed and features are chosen before the model is built. Evaluation metrics are employed, such as F1-score, recall, accuracy, and precision. The SVM model performs better than the others, with an impressive accuracy of 91.67% [10]. Lilhore et al.(2022) investigate the application of Machine Learning (ML) in predicting heart diseases and locomotor disorders using a UCI dataset with 303 rows and 76 properties. Focusing on 14 selected properties, the study employs various ML methods, including Naive Bayes, SVM, Logistic regression, Decision Tree Classifier, Random Forest, and K-Nearest Neighbor (KNN). The isolation forest approach is applied to enhance precision. Experimental results highlight the effectiveness of KNN with eight neighbors, outperforming other methods in terms of sensitivity, precision, accuracy, and F1-score compared to Naive Bayes, SVM (Linear Kernel), Decision Tree Classifier with 4 or 18 features, and Random Forest classifiers [11].

III. PROPOSED METHODOLOGY

The suggested approach makes use of cutting-edge machine learning methods to predict the possibilities of Cardiovascular problems. A Decision Tree classifier, a potent algorithm renowned for its interpretability and effectiveness in identifying complicated relationships within datasets, is used to generate a strong solution. A large and diverse dataset including age, sex, thalassemia, etc.. is used to train the Decision Tree model. This feature set guarantees a comprehensive analysis of several factors that may impact heart health, offering a strong basis for precise forecasts. In the domain of heart disease prediction, the Decision Tree classifier was selected because of its capacity to manage both numerical and categorical data with effectiveness. It is expected to produce findings that are clear and understandable.

A. DATASET

Fourteen unique features make up the dataset used to forecast heart disease, and each one provides essential data for precise predictive modeling and a clear definition of each one

is depicted in Table I. These characteristics provide a broad base for thorough machine learning analysis by encompassing a variety of demographic, clinical, and physiological factors. The value of these characteristics is found in their combined capacity to capture many aspects of a person's health, offering a comprehensive viewpoint that helps with the accurate prognosis of heart disease. The abundance and variety of these characteristics enable the machine learning model to identify minute patterns and connections, which eventually improves the prediction's precision and dependability.

TABLE I
DATASET DESCRIPTION

age	Person's age
sex	Person's gender
cp	A specific kind of chest discomfort
trestbps	Blood pressure at rest
chol	Blood cholestrol levels
restecg	Resting electrocardiographic data
fbs	Blood sugar levels when fasting
thalach	Highest heart rate possible while exercising
exang	Angina brought by activity
oldpeak	Exercise induced ST depression
slope	Peak exercise ST segement's slope
ca	Quantity of significant vessels shown by fluoroscopy
thal	Stands for Thalassemia
target	Indicates if Cardiac disease is present or not by '0' or '1'

B. ANALYSING THE DATASET

Comprehending the target variable's distribution, shown in Figure 1, is essential for assessing any prediction model's efficacy. The target variable in this dataset has binary values of 1 or 0, denoting the existence or absence of cardiac disease. Important information on the prevalence of heart disease in the population under study is provided by the target distribution analysis.

The dataset is examined, and it is found that the distribution is fairly balanced, with similar numbers of examples for each of the two classes. When training a machine learning model, this balanced distribution is helpful since it makes sure the model sees a variety of instances for both positive and negative scenarios. The dataset's balance helps the model generalize well to previously undiscovered data, improving its predictive power.

Building a strong heart disease prediction model requires a solid understanding of the target distribution. The learning process is aided by a balanced dataset, which helps the model identify patterns and relationships that help it make correct predictions. This analysis emphasizes how crucial it is to take class distribution into account when analyzing the model's performance measures and shows how well-suited the dataset is for developing a trustworthy predictive model for heart disease.

Knowing the age-specific distribution of heart disease, depicted in Figure 2, is essential to comprehending the risk factors and prevalence of cardiovascular health. Age is a key demographic characteristic in the dataset that offers important insights on how the prevalence of heart disease differs in

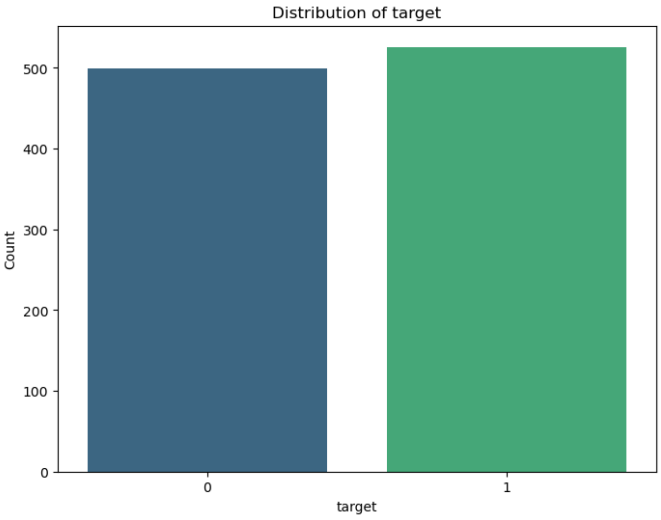

Fig. 1. Classwise Distribution of Dataset

different age groups. The analysis that follows looks at the age-specific distribution of heart disease and highlights any possible trends or patterns. It is clear from looking at the dataset that a range of ages are included, which makes it possible to conduct a thorough investigation of the factors associated with aging and heart disease. Statistical variables like mean age, age range, and age-specific disease prevalence can be generated to show this distribution. Furthermore, graphical displays like age-specific bar graphs or histograms can clearly convey the distribution patterns. The dataset's analysis shows how being older may alter one's risk of heart disease. This detailed knowledge is essential for determining risk factors particular to an individual's age, developing preventative strategies, and developing personalized healthcare solutions. These kinds of discoveries can make a big difference in public health programs that try to address and lessen the effects of heart disease in particular age groups.

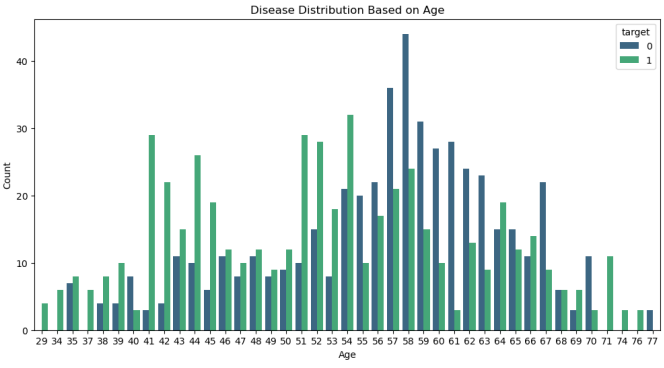

Fig. 2. Disease Distribution based on Age

C. DECISION TREE

With an abundance of qualities that set them apart, Decision Trees prove to be a powerful ally in our quest for precise

heart disease prediction. Decision trees are elegant because they reduce complicated decision-making procedures to a set of simple, understandable options. In addition to promoting a deeper comprehension of the model's predictions, this openness fosters acceptance and confidence within the medical community.

The use of metrics like Gini Index, which directs the selection of the most discriminative characteristics at each decision point, is a crucial component of decision trees. This flexible methodology allows the model to capture complex patterns in the dataset, which is important in the complex field of medical diagnosis.

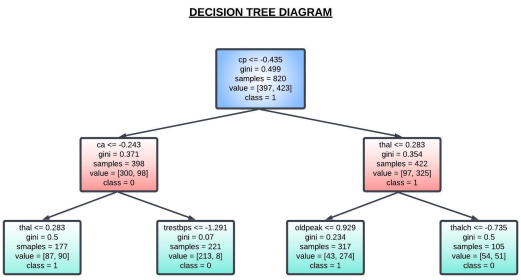

Fig. 3. Decision Tree Classifier

Decision trees' principal advantage is their organizational structure. The dataset is systematically divided into subgroups by the model, which then builds a decision tree with nodes signifying important choices and leaves representing definitive predictions, which is shown in Figure 3. This structure provides a concrete and understandable framework for decision-making, imitating the diagnostic procedure used by medical professionals.

A decision tree's workflow is similar to a diagnostic process. Figure 4 tells us how the model starts at the root and iterates over the features, picking the most illuminating ones to generate decision points. This process's iterative structure and flexibility to various data distributions guarantee the model's resilience in a range of healthcare contexts.

The Decision Tree model demonstrated its efficacy in our research undertakings with an astounding 99% accuracy rate in heart disease prognosis. More than just being accurate, its interpretability turns into a critical advantage that enables healthcare professionals to identify the variables affecting forecasts and promotes teamwork in the delivery of patient care.

To put it simply, the Decision Tree is more than just a prediction model; it is a partner in the complex dance of healthcare diagnostics, providing a balance of precision, interpretability, and flexibility that is essential to the quest for better patient outcomes.

D. EVALUATION CRITERIA

The evaluation metrics offer a thorough analysis of how well the Decision Tree model predicts cardiac disease. Excellent recall, precision, and F1-score values are shown in the classification report for both positive and negative classes. The model has great accuracy in properly recognizing instances of no heart disease and heart disease, with precision values of 0.97 and 1.00 for class 0 and class 1, respectively. The model's recall values of 1.00 for class 0 and 0.97 for class 1 show that it can accurately identify both the majority of heart disease cases and all cases of no heart disease.

The harmonic mean of precision and recall, or F1-score, is 0.99 for both classes, indicating that the model performs fairly in both categories. The model's overall accuracy of 0.99 indicates that it is capable of producing accurate predictions for the whole dataset.

The remarkable AUC value of 0.99 on the ROC curve, shown in Figure 5, further highlights the model's capacity to distinguish between positive and negative cases across a range of classification thresholds. The model's superiority in retaining high precision and recall is reinforced by the PR curve's AUC value of 0.99, depicted in Figure 6, which is especially important in scenarios with imbalanced datasets.

A solid basis for the Decision Tree model's use in heart disease prediction is provided by the evaluation criteria taken together, which show the model's resilience. High precision, recall, and AUC values highlight how well the model works to differentiate between people with and without heart disease.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

Additionally, we examine possible overfitting and underfitting of our Decision Tree model in addition to accuracy measures. Generalization to new data can be hampered by overfitting, a condition in which the model learns the training data too thoroughly. On the other hand, underfitting happens when the training data's underlying patterns are missed by the model.

The 100 % training accuracy indicates that our model has assiduously mastered the training set. Nonetheless, the 99 % validation accuracy suggests that the model fits unknown data well, allaying worries about overfitting, depicted in Figure 7. This is further corroborated by the cross-validation scores, which demonstrate consistently good accuracies across various dataset subsets.

Our model has struck a desirable equilibrium, as evidenced by the high cross-validation scores and the insignificant difference between training and validation accuracies. It shows a strong capacity for learning from the training data and generalizing to new situations. These results strengthen the model's potential for use in real-world healthcare scenarios by demonstrating the model's reliability and adaptability to a variety of datasets.

The model's confusion matrix in Figure 8 shows encouraging outcomes that point to good predictive performance. The model shows good accuracy in detecting positive and negative situations, properly classifying 102 occurrences as real positives and 100 as true negatives. The minimal amount of false positives (3) and false negatives (0) highlights the accuracy and dependability of the model in reducing misclassifications. This implies that the machine learning model strikes an impressive

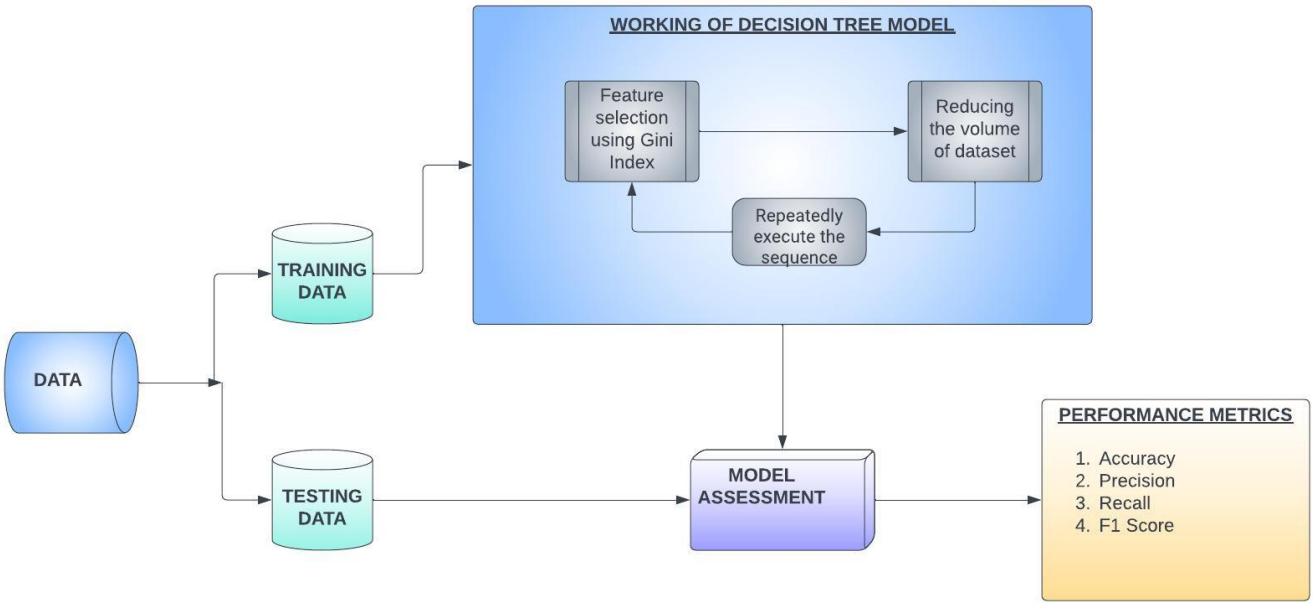

Fig. 4. Work Flow of Cardiovascular Disease Prediction model based Decision Tree

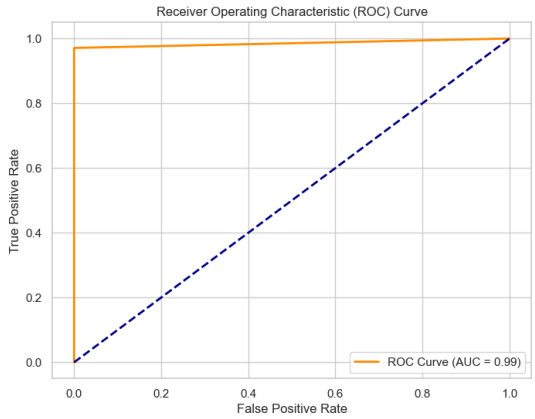

Fig. 5. Roc Curve obtained in the Proposed Model

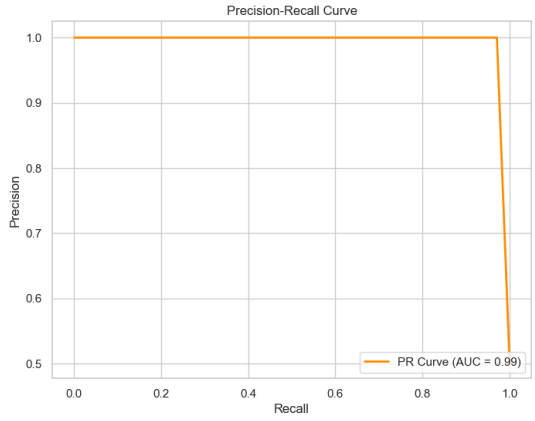

Fig. 6. Precision- Recall curve obtained in the Proposed Model

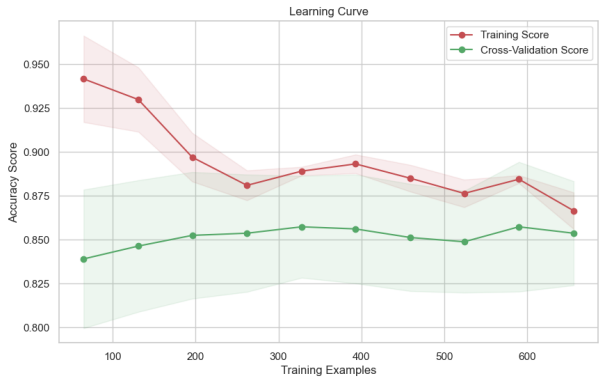

Fig. 7. Investigation results of Overfitting and Underfitting in the Proposed Model

balance between sensitivity and specificity, especially when it comes to heart disease prediction, which adds to its overall diagnostic efficacy.

A visual depiction of the connections between the various attributes in the dataset is offered by scatter plots in Figure 9. Key characteristics like age, cholesterol (chol), maximum heart rate attained (thalach), exercise-induced ST depression (oldpeak), and resting blood pressure (trestbps) can all be examined as scatter plots to get important insights into any correlations or trends.

Any patterns or correlations between age and cholesterol levels can be shown with the scatter plot of age vs cholesterol (chol). Similar scatter plots show possible relationships or differences in these characteristics between various age groups for age versus maximal heart rate (thalach), exercise-induced

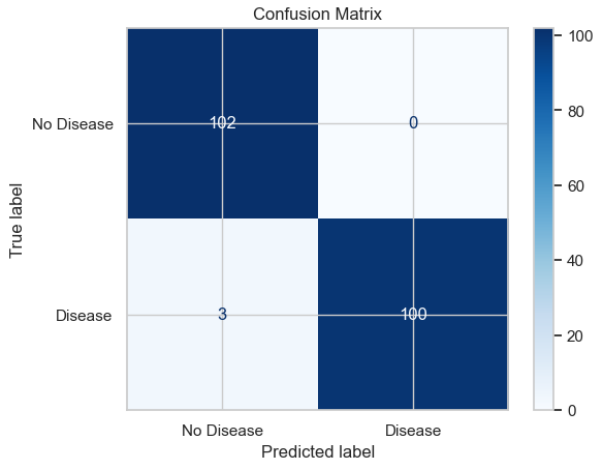

Fig. 8. Confusion Matrix obtained in the Proposed Model

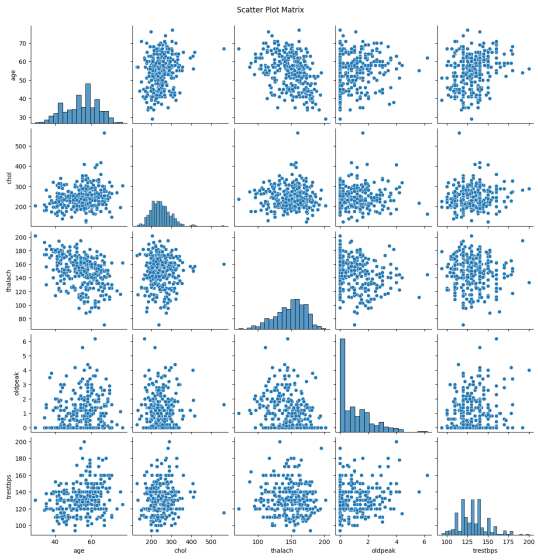

Fig. 9. Distribution of Main features

ST depression (oldpeak), and resting blood pressure (trestbps).

By examining these scatter plots, important details can be found, such as whether a certain attribute tends to rise or fall with age, or whether any patterns stand out that might be related to cardiovascular health. The process of visualizing these correlations facilitates the identification of potential risk factors for heart disease and leads to a more thorough comprehension of the dataset.

We thoroughly evaluated the effectiveness of various machine learning methods, such as Decision Tree, Logistic Regression, Support Vector Machine (SVM), k-Nearest Neighbors (KNN), and Naive Bayes, in our investigation of cardiovascular disease prediction models. With an astounding accuracy rate of 99%, the Decision Tree algorithm was the clear winner among this group of algorithms. The Decision Tree model is unique not only because of its great accuracy but also because of its remarkable resistance to both overfitting and underfitting, which guarantees stable adaptation to new datasets. While all four models—Logistic Regression, SVM,

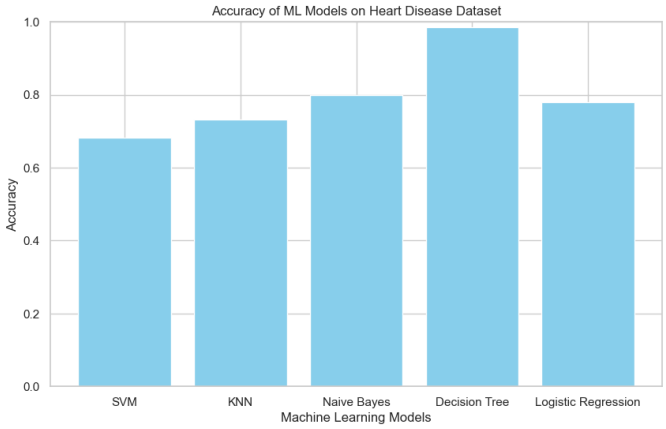

Fig. 10. Comparison of Proposed Models with other Models

KNN, and Naive Bayes—performed admirably, none was able to match the Decision Tree model's accuracy and generalization capacity which is shown in Figure 10. This result confirms that the Decision Tree algorithm is the best option for our particular goal of predicting cardiovascular illness. Because of its dependability and efficiency, it can be used as a powerful tool for cardiovascular health early detection and intervention. Its exceptional performance was a result of the careful feature selection and model optimization, which made it an invaluable tool in the continuous effort to improve healthcare outcomes.

CONCLUSION

The work uses a decision tree method to predict cardiovascular disease in humans, outperforming previous machine learning algorithms - achieving 99% accuracy. It can be used to identify people at risk of heart disease. The decision tree system and excellent performance show its potential as a useful tool for early diagnosis and preventive measures. The materials chosen for this project were critical to modelling the process and improving efficiency in addition to remarkable accuracy. The predictive power of the model was improved by defining and organizing the most important variables in the data. The research and its implications are significant and provide a good opportunity for prevention and early detection of cardiovascular disease. This work provides a robust and reliable predictive model that should improve health outcomes, contributing to ongoing efforts to combat heart disease. Further research may explore partnerships, larger datasets and practical applications to fully understand the potential implications of this predictive model for cardiovascular health.

REFERENCES

[1] Sivakannan Subramani , Neeraj Varshney , M. Vijay Anand , Manzoore Elahi M. Soudagar , Lamya Ahmed Al-keridis , Tarun Kumar Upadhyay , Nawaf Alshammari , Mohd Saeed , Kumaran Subramanian , Krishnan Anbarasu and Karunakaran Rohini "Cardiovascular diseases prediction by machine learning incorporation with deep learning " Volume 10 - 2023

[2] J.Gabriel Ott, Yannik Schaubelt, Juan Miguel Lopez Alcaraz, Wilhelm Haverkamp, Nils Strodthoff "Uncovering ECG Changes during Healthy Aging using Explainable AI" Submitted on 11 Oct 2023

[3] Bonaventure F. P. Dossou, Yenoukoume S. K. Gbenou, Miglanche Ghomsi Nono "Pretrained vision models for predicting high-risk breast cancer stage" Published as a conference paper at ICLR 2023

[4] Nadiah A. Baghdadi , Sally Mohammed Farghaly Abdelaliem , Amer Malki , Ibrahim Gad , Ashraf Ewis and Elsayed Atlam "Advanced machine learning techniques for cardiovascular disease early detection and diagnosis " volume 10, Article number: 144 (2023)

[5] Bruno Machado Pacheco, Victor Hugo Rocha de Oliveira, Augusto Braga Fernandes Antunes, Saulo Domingos de Souza Pedro, Danilo Silva "Does pre-training on brain-related tasks results in better deep-learning-based brain age biomarkers?" 11 Jul 2023

[6] P. Chinnasamy , S. Arun Kumar , V. Navya , K. Lakshmi Priya , Siva Sruthi Boddu "Machine learning based cardiovascular disease prediction" Volume 64, Part 1, 2022

[7] Fakir, Youssef, and Naoum Abdelmotalib. "Analysis of Decision Tree Algorithms for Diabetes Prediction." In International Conference on Business Intelligence, pp. 197-205. Cham: Springer International Publishing, 2022.

[8] El-Shafiey, Mohamed G., Ahmed Hagag, El-Sayed A. El-Dahshan, and Manal A. Ismail. "A hybrid GA and PSO optimized approach for heart-disease prediction based on random forest." Multimedia Tools and Applications 81, no. 13 (2022): 18155-18179.

[9] Poojitha, T., and R. Mahaveerakannan. "Classification and Prediction of Heart Disease using Novel Random Forest Algorithm by Comparing Logistic Regression for Obtaining Better Accuracy." Cardiometry 25 (2022): 1538-1545.

[10] Boukhatem, Chaimaa, Heba Yahia Youssef, and Ali Bou Nassif. "Heart disease prediction using machine learning." In 2022 Advances in Science and Engineering Technology International Conferences (ASET), pp. 1-6. IEEE, 2022.

[11] Ramesh, T. R., Umesh Kumar Lilhore, M. Poongodi, Sarita Simaiya, Amandeep Kaur, and Mounir Hamdi. "Predictive analysis of heart diseases with machine learning approaches." Malaysian Journal of Computer Science (2022): 132-148.