IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134183

PRECISE HEART: HEART DISEASE PREDICTION USING MACHINE LEARNING

Mohankumar N[1], Kavinandhan B[2], Pranav R[3], Vinu Prasanth MJ[4]

Assistant Professor, Department of ECE, Amrita School of Engineering, Coimbatore, India[1]

Student, Department of ECE, Amrita School of Engineering, Coimbatore, India[2-4]

Abstract: In most cases, heart disease diagnosis depends on a complex combination of clinical and pathological data. Because of this complexity, there is a significant amount of interest among clinical professionals and researchers regarding efficient and accurate heart disease prediction. In this paper, we develop a heart disease prediction system that can assist medical professionals in predicting heart disease status based on the clinical data of patients. The system will consist of multiple features, including an input clinical data section, ROC curve display section, and prediction performance display section (execute time, accuracy, sensitivity, specificity, and predict result). The paper also discusses the pre-processing methods, classifier performances, and evaluation metrics. We have investigated the accuracy levels of various machine learning techniques such as Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Naive Bayes, and Decision Trees (DT). In the result section, the visualized data shows that the prediction is accurate. The system developed in this study proves to be a novel approach that can be used in the classification of heart disease.

Keywords: Statistical Description and Dispersion, Correlation, Feature Analysis, Classification, K-Nearest Neighbor, Decision Tree, Support Vector Machines, Naive Bayes

I. INTRODUCTION

The work proposed in this paper focuses mainly on various data mining practices that are employed in heart disease prediction. The human heart is the principal part of the human body. It regulates blood flow throughout our bodies, any irregularity to the heart can cause distress in other parts of the body. In today's contemporary world, heart disease is one of the primary reasons for the occurrence of most deaths. It may occur due to an unhealthy lifestyle, smoking, alcohol, and high intake of fat which may cause hypertension [1].

According to the World Health Organization, more than 10 million die yearly from heart diseases. A healthy lifestyle and the earliest detection are the only ways to prevent heart-related diseases. The main challenge in today's healthcare is the provision of quality services and effective accurate diagnoses.[2] Even though heart diseases are found to be more prominent in recent years, they are also the ones that can be controlled and managed effectively. The whole accuracy in the management of disease lies in the proper time of detection of that disease. The proposed work attempts to detect these heart diseases at an early stage to avoid disastrous consequences. Records of a large set of medical data created by medical experts are available for analyzing and extracting valuable knowledge from it.

Data mining is a multidisciplinary field, drawing work from areas including database technology, machine learning, statistics, pattern recognition, information retrieval, neural networks, knowledge-based systems, artificial intelligence, high-performance computing, and data visualization. Mostly the medical database consists of discrete information, thereby decision-making becomes a complex task. A data analysis system that does not handle large amounts of data should be more appropriately categorized as a machine learning system, a statistical data analysis tool, or an experimental system prototype.

A system that can only perform data or information retrieval, including finding aggregate values, or that performs deductive query answering in large databases should be more appropriately categorized as a database system, an information retrieval system, or a deductive database system. In the medical field, machine learning can be used to diagnose, detect, and predict various diseases. The main goal of this paper is to provide a tool for doctors to detect heart disease at an early stage. This will help provide effective treatment to patients and avoid severe consequences. This paper presents a details performance analysis using various machine learning techniques, K-means, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naïve Bayes, and Decision Tree [3].

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134183

II. MOTIVATION

The main objective of doing this research is to present a heart disease prediction model for the prediction of the occurrence of heart disease. Further, this research work is aimed toward identifying the best classification algorithm for identifying the possibility of heart disease in a patient. It is justified by performing a comparative study and analysis using four classification algorithms: Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naïve Bayes and Decision Tree. These are used at different levels of evaluation. Although these are commonly used machine learning algorithms, heart disease prediction is a vital task involving the highest possible accuracy. Hence, the four algorithms are evaluated at numerous levels and types of evaluation strategies.

This will provide researchers and medical practitioners to establish a better. The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but they are either expensive or are not efficient to calculate the chance of heart disease in humans. Early detection of cardiac diseases can decrease the mortality rate and overall complications.

However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient 24 hours, since it requires more sapience, time and expertise. With a good amount of data in today's world, we can use various machine learning algorithms to analyze the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data.

III. LITERATURE SURVEY

The intersection of medical science and machine learning has spurred significant advancements in recent years, particularly in the realm of heart disease prediction. Researchers have explored various methodologies and models to enhance the accuracy and efficiency of prediction systems, aiming to provide timely and reliable clinical decision support. Several studies have delved into the application of machine learning techniques for real-time heart disease prediction. Chintan M. Bhatt, Parth Patel et al. (2023) contributed to this field by developing a model tailored for real-time prediction, demonstrating the potential of machine learning in augmenting traditional diagnostic approaches [4]. Similarly, Harshit Jindal, Sarthak Agarwal, Rishabh Khera Jain and Preeti Nagrath (2020) proposed an effective prediction model, highlighting its utility in clinical decision support systems.

Comparative analyses have also been conducted to evaluate the performance of different machine learning techniques in heart disease prediction [5]. Tulika Lodh, Anirban Dey, Naorem Rinita, Sunil Kumar, Subodh Kumar et al. (2021) conducted a comprehensive study comparing various methodologies, shedding light on the strengths and limitations of different approaches. Such analyses are crucial for identifying the most suitable techniques for specific contexts and optimizing prediction accuracy.

Furthermore, historical evaluations have provided insights into the evolution of machine learning techniques for heart disease prediction [6]. Keshav Srivastava and Dilip Kumar Choubey (2020) evaluated the performance of different methodologies, laying the groundwork for subsequent research endeavors and showcasing the progression of predictive models over time [7]. Collectively, these studies underscore the growing significance of machine learning in augmenting heart disease prediction. By leveraging advanced algorithms and vast datasets, researchers continue to refine prediction models, ultimately contributing to improved patient care and outcomes

IV. METHODOLOGY

A. Existing System

Heart disease is even being highlighted as a silent killer which leads to the death of a person without obvious symptoms. The nature of the disease is the cause of growing anxiety about the disease & its consequences. Hence continued efforts are being done to predict the possibility of this deadly disease in prior. So various tools & techniques are regularly being experimented with to suit the present-day health needs. Machine Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them under suitable headings & finally analyzing to extract the desired data we can conclude. This technique can be very well adapted to the do the prediction of heart disease. As the well-known quote says "Prevention is better than cure", early prediction & its control can be helpful to prevent & decrease the death rates due to heart disease.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134183

B. Proposed System

The proposed work predicts heart disease by exploring the above-mentioned four classification algorithms and carrying out performance analysis. The objective of this study is to effectively predict if the patient suffers from heart disease. The health professional enters the input values from the patient's health report. The data is fed into the model which predicts the probability of having heart disease.

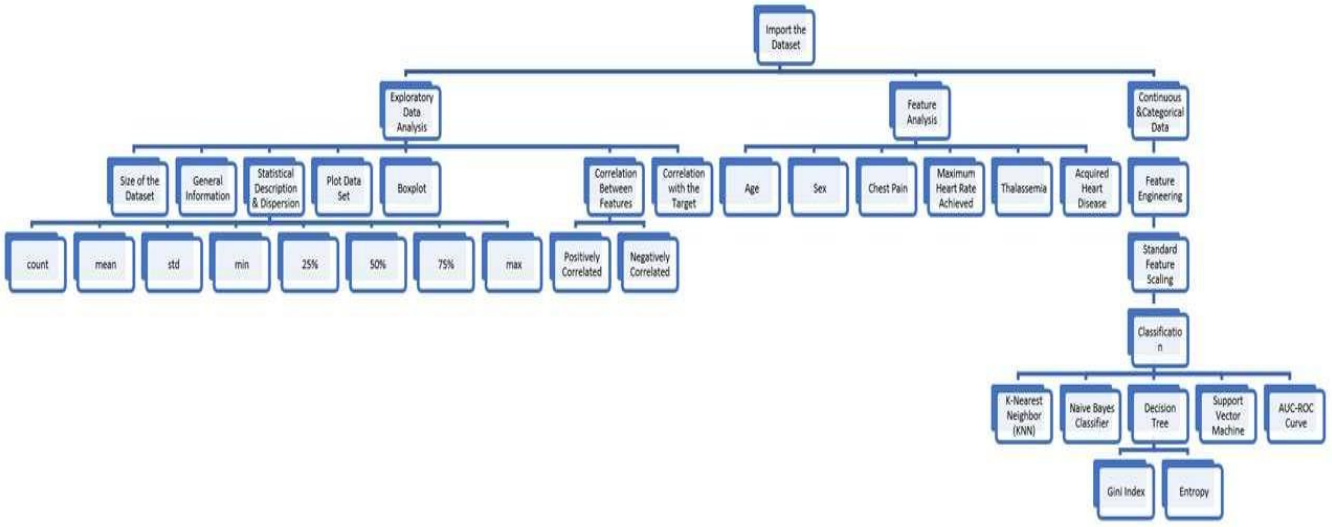

Fig. 1. Generic model to predict heart disease

Initially, we collect a dataset for our heart disease prediction system. After the collection of the dataset, we split the dataset into training data and testing data. The training dataset is used for prediction model learning and testing data is used for evaluating the prediction model. For this project, 80% of training data is used and 20% of data is used for testing. The dataset used for this project is Heart Disease UCI. The dataset consists of 76 attributes; out of which, 14 attributes are used for the system.

C. Attribute Information

The dataset is a combination of 4 different databases, but the primary one is the UCI Cleveland dataset. This database consists of a total of 76 attributes but all published experiments refer to using a subset of only 14 features.[6] Therefore, we have used the already processed UCI Cleveland dataset available on the Kaggle website for our analysis.

Table-1. Attribute Information

Serial Number	Attribute	Distinct Values ofAttribute
1	Age (in Years)	NIL
2	Sex	Female (0)
		Male (1)
3	Chest Pain	Asymptomatic (0)
		Nonanginal (1)
		Nontypical (2)
		Typical (3)
4	Resting Blood Pressure (mm Hg onadmission to the hospital)	NIL

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134183

5	Serum CholesterolMeasurement (mg/dl)	NIL
6	Fasting Blood Sugar > 120 mg/dl	False (0)
		True (1)
7	Resting Electrocardiographic Results	Showing probable or definite left ventricularhypertrophy by Estes' criteria (0)
		Normal (1)
		Having ST-T wave abnormality (T wave inversions and/or ST elevation or depressionof > 0.05 mV) (2)
8	Maximum HeartRate Achieved	NIL
9	Exercise-InducedAngina	No (0)
		Yes (1)
10	Old Peak – ST depression inducedby exercise relativeto rest ('ST' relatesto positions on the ECG plot)	NIL
11	The slope of the peak exercise STsegment	Down Sloping (1)
		Flat (2)
		Up Sloping (3)
12	Number of major vessels coloured byfluoroscopy	0
		1
		2
		3
13	A blood disorder called thalassemia	Dropped from the dataset previously (0 –NA)
		No blood flow in someparts of the heart (1 – fixed)
		Normal blood flow (2 –normal)
		A blood flow is observed but it is not normal (3 – reversible)
14	Acquired Heart Disease (AHD), Output Class	Normal (0 – No)
		Heart Disease (1 – Yes)

D. Exploratory Data Analysis

Correlation analysis is a cornerstone of data analysis, especially in predictive modeling endeavors like heart disease prediction. It serves as a compass, guiding researchers through the intricate web of relationships among various variables within a dataset. By scrutinizing these relationships, correlation analysis unveils patterns and dependencies that underlie the data, offering invaluable insights into how different features interact and influence one another. This understanding is pivotal for crafting accurate predictive models, as it empowers researchers to discern which variables hold the most predictive power and how they interplay in predicting heart disease risk. Through correlation analysis, researchers gain a nuanced understanding of the data landscape, paving the way for more informed modeling decisions and ultimately enhancing the predictive performance of the models.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134183

Fig. 4. Correlation Between Various Features

Attribute or feature selection includes the selection of appropriate attributes for the prediction system. This is used to increase the efficiency of the system.

Various attributes of the patient like gender, chest pain type, fasting blood pressure,serum, cholesterol, etc. are selected for the prediction. TheCorrelation matrix is used for attribute selection for this model.

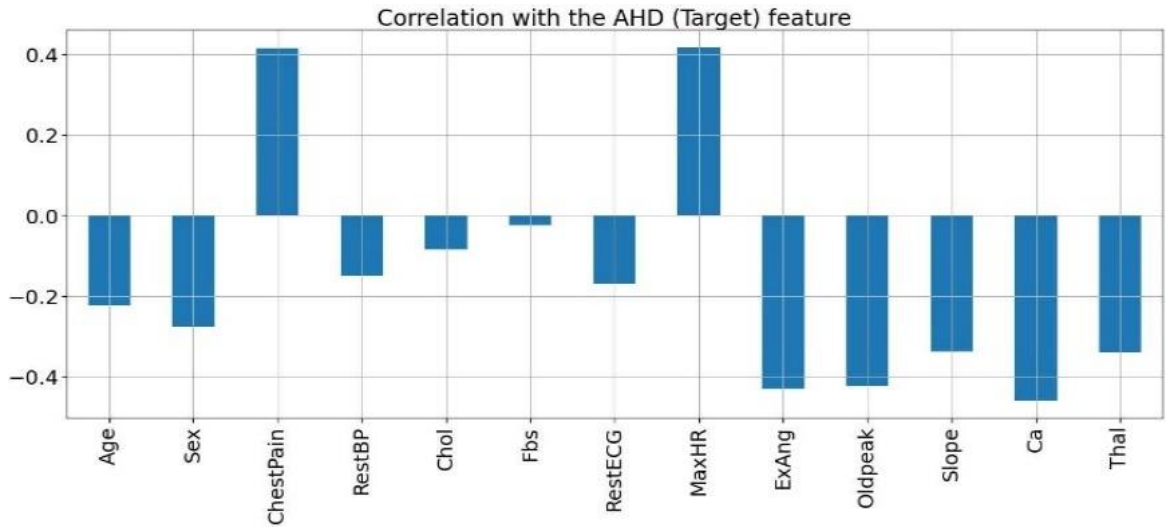

Fig. 5. Correlation with the Acquired Heart Disease(Target) feature

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134183

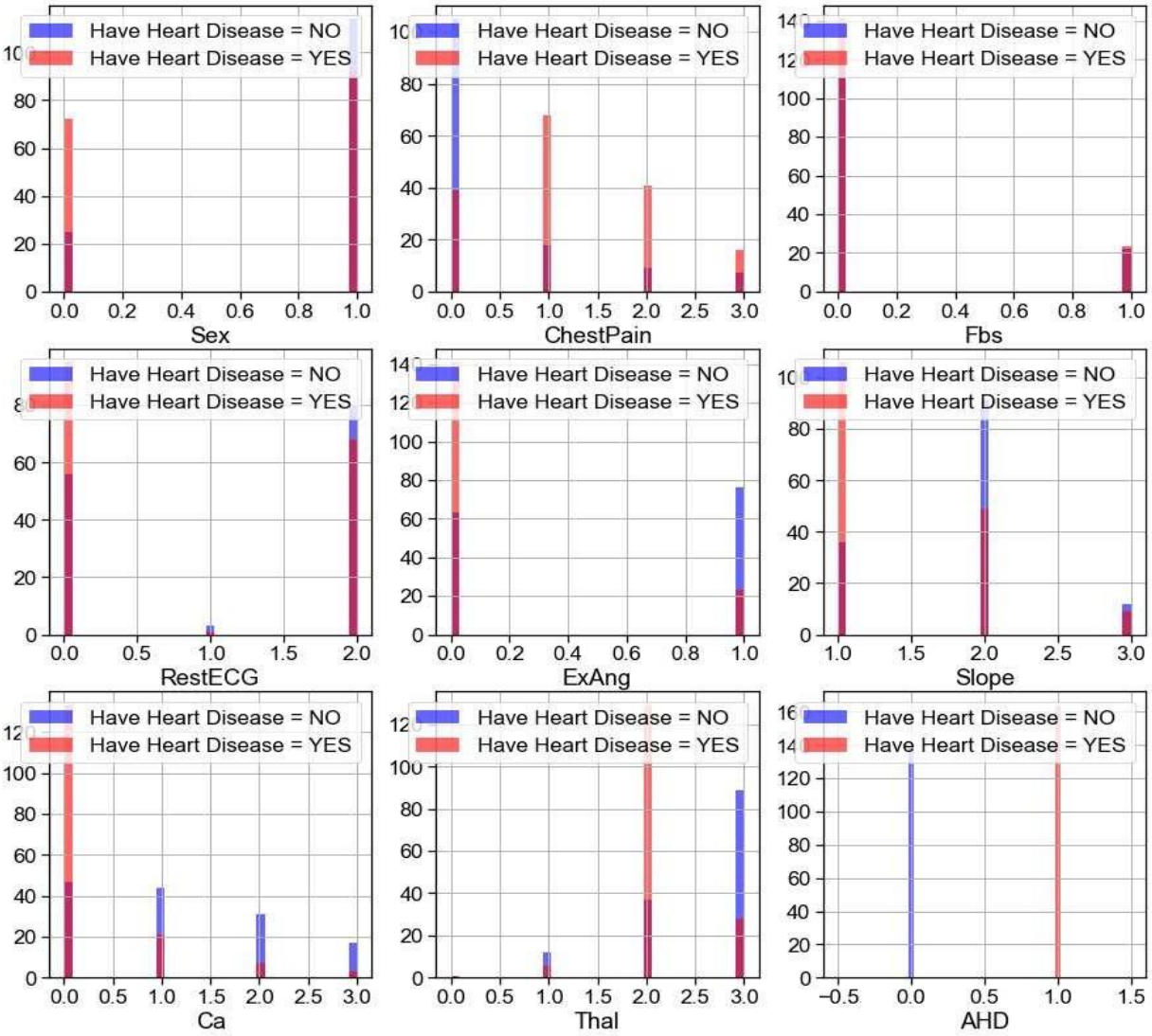

Fig. 6. Acquired Heart Disease based on Categorical Data

Chest Pain – People with chest pain equal to 1, 2, and 3 are more likely to have heart disease than people with chest pain equal to 0.

Resting Electrocardiographic Results – People with a value of 0 (showing probable or definite left ventricular hypertrophy by Estes' criteria, which can range from mild symptoms to severe problems) are more likely to have heart disease.

Exercise-Induced Angina – People with a value of 0 (No) have heart disease more than people with a value of 1 (Yes).

Slope – People with a slope value equal to 1 (Down-sloping - Signs of Unhealthy Heart) are more likely to have heart disease than people with a slope value equal to 2 (Up-sloping - Better Heart Rate with Exercise) or 3 (Flat - Minimal Change, Typical Healthy Heart).

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134183

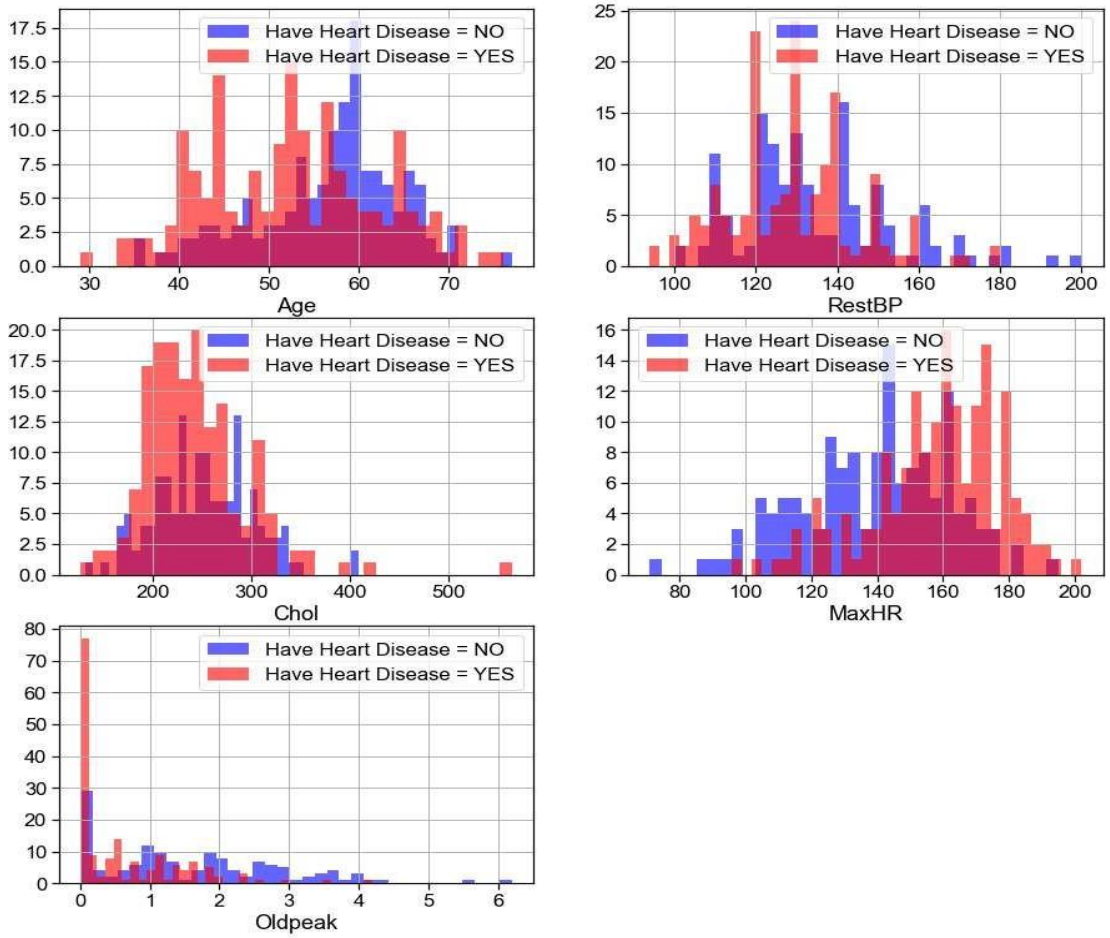

Fig. 6. Acquired Heart Disease based on Important Features

Age – Elderly people (>50 years) are more likely to have heart disease.

Resting Blood Pressure – Anything between 120-140 (mm Hg on admission to the hospital) is typically a cause for concern.

Serum Cholesterol Measurement – Anything between 200-300 (mg/dl) is typically a cause for concern.

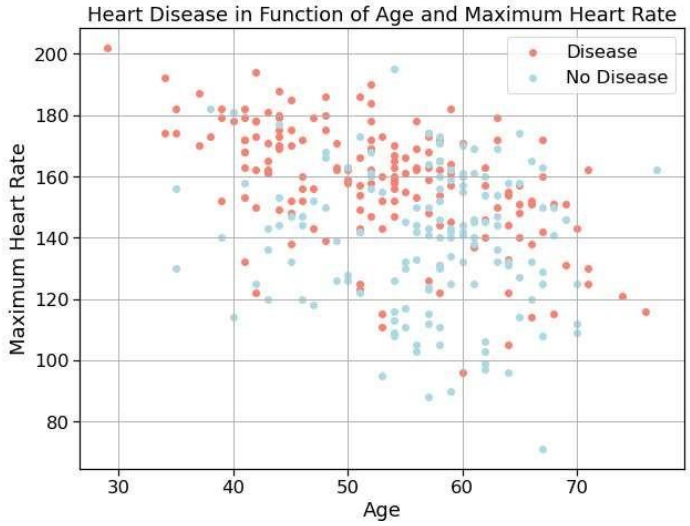

Fig. 6. Heart Disease in Function of Age and MaximumHeart Rate

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134183

E. Classification

	Naive Bayes Classifier	Decision Tree Using Gini Index	Decision Tree Using Entropy	Support Vector Machine (SVM)	K-Nearest Neighbour (KNN)
Confusion Matrix	28 4 3 26	27 5 27 22	26 6 3 26	26 6 5 24	28 4 2 27
Accuracy	88.52459016	80.32786885	85.24590164	81.96721311	90.16393443
Report					
0	0.9 / 0.88 / 0.89 / 32	0.79 / 0.84 / 0.82 / 32	0.9 / 0.81 / 0.85 / 32	0.84 / 0.81 / 0.83 / 32	0.93 / 0.88 / 0.9 / 32
1	0.87 / 0.9 / 0.88 / 29	0.81 / 0.76 / 0.79 / 29	0.81 / 0.9 / 0.85 / 29	0.8 / 0.83 / 0.81 / 29	0.87 / 0.93 / 0.9 / 29
Accuracy	0.89 / 61	0.8 / 61	0.85 / 61	0.82 / 61	0.9 / 61
Macro Average	0.88 / 0.89 / 0.89 / 61	0.8 / 0.8 / 0.8 / 61	0.85 / 0.85 / 0.85 / 61	0.82 / 0.82 / 0.82 / 61	0.9 / 0.9 / 0.9 / 61
Weighted Average	0.89 / 0.89 / 0.89 / 61	0.8 / 0.8 / 0.8 / 61	0.86 / 0.85 / 0.85 / 61	0.82 / 0.82 / 0.82 / 61	0.9 / 0.9 / 0.9 / 61
AUC Score	91.91810345	89.92456897	88.41594828	89.54741379	93.53448276

Fig. 7. Machine Learning Techniques with Accuracy

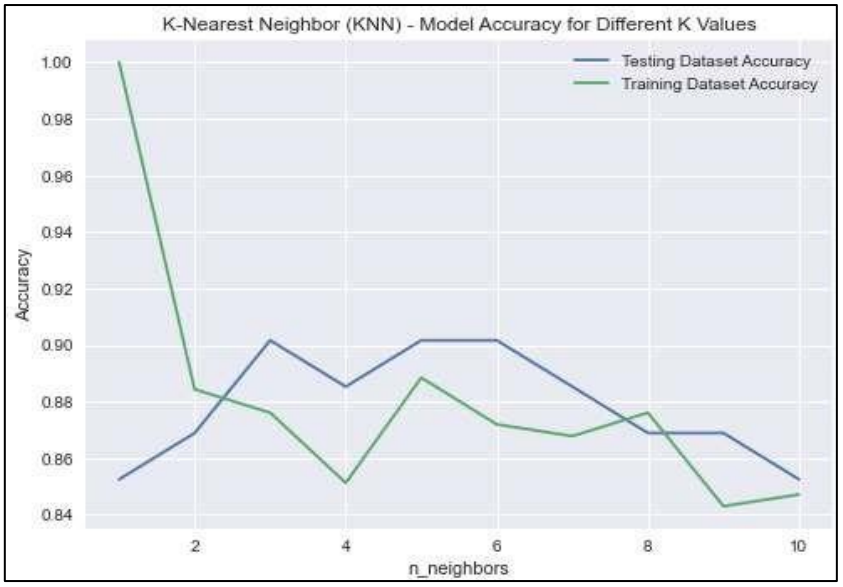

Fig. 8. K-Nearest Neighbor (KNN) - Model Accuracy for Different K Values

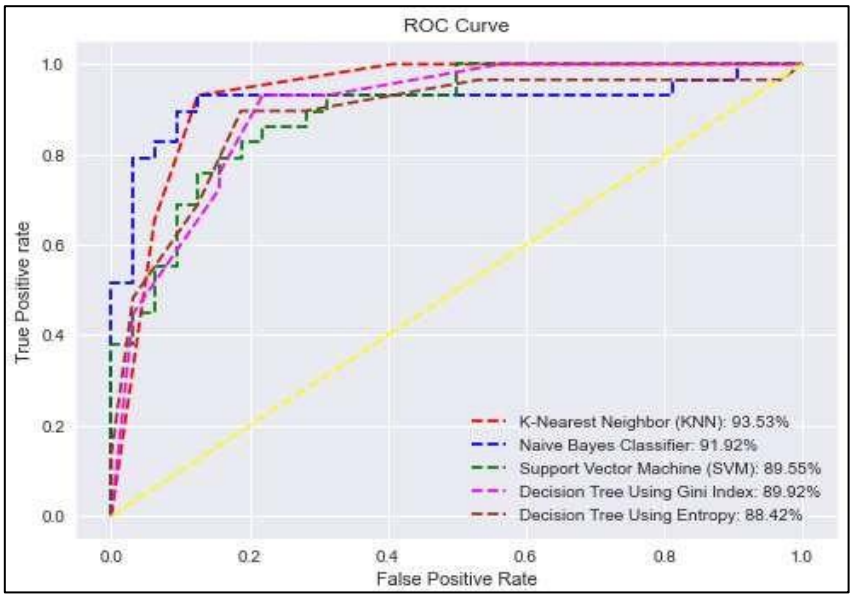

Fig. 9. ROC Curve

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134183

V. CONCLUSION

With the increasing number of deaths due to heart diseases, it has almost become increasingly mandatory to develop a proficient system to predict heart diseases effectively and accurately. This study compares the accuracy score of K-Nearest Neighbor (KNN), Naïve Bayes Classifier, Support Vector Machine, and Decision Tree algorithms for predicting heart disease using the UCI machine learning repository dataset. The result of this study indicates that the K-Nearest Neighbor (KNN) algorithm is the most efficient algorithm with an accuracy score of 93.53% for the prediction of heart disease. In the future, the work can be enhanced by developing a web application based on the K-Nearest Neighbor (KNN) as well as using a larger dataset as compared to the one used in this analysis, which will help to provide better results and help health professionals in predicting the heart disease effectively and efficiently

REFERENCES

[1]. T. Nagamani, S. Logeswari, B. Gomathy, Heart Disease Prediction using Data Mining with Mapreduce Algorithm, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-3, January 2019.

[2]. Avinash Golande, Pavan Kumar T, Heart Disease Prediction Using Effective Machine Learning Techniques, International Journal of Recent Technology and Engineering, Vol 8, pp.944- 950, 2019.

[3]. Fahd Saleh Alotaibi, Implementation of Machine Learning Model to Predict Heart Failure Disease, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6, 2019.

[4]. Researchers Chintan M. Bhatt and Parth Patel et al. "Effective Heart Disease Prediction Using Machine Learning techniques". Multidisciplinary Digital Publishing Institute (MDPI), 2023

[5]. Harshit Jindal, Sarthak Agarwal, Rishabh Khera Jain and Preeti Nagrath proposed "Heart disease prediction using machine learning algorithms" 2020

[6]. Tulika Lodh, Anirban Dey, Naorem Rinita, Sunil Kumar, Subodh Kumar et al, proposed a paper "Analysis of Heart Disease Prediction using Machine Learning Techniques". International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET) 2021

[7]. Keshav Srivastava and Dilip Kumar Choubey published a paper on "Heart Disease Prediction using Machine Learning and Data Mining" in International Journal of Recent Technology and Engineering (IJRTE), 2020