International Journal of Scientific Research & Engineering Trends
Volume 10, Issue 6, Nov-Dec-2024, ISSN (Online): 2395-566X

Heart Disease Detection Using Machine Learning

Assistant Professor Ms. Pragati, Mr. Shivam Chawla, Mr. Yash Mittal, Mr. Shivam Mishra
CSE Department, HMR Institute of Technology and Management, GGSIPU, Delhi

Abstract- Cardiovascular diseases (CVDs) are a leading cause of death worldwide, posing a significant health threat not only in India but across the globe. This highlights the critical need for a dependable, precise, and accessible system to diagnose such conditions promptly, enabling timely treatment. Machine learning algorithms have become invaluable tools in healthcare, automating the analysis of extensive and complex datasets. Recent studies demonstrate that various machine learning techniques can aid healthcare professionals in diagnosing heart-related conditions. The heart, second only to the brain in importance, plays a vital role in circulating blood throughout the body. Predicting heart disease occurrence is thus essential in the medical field. Data analytics enhances the prediction accuracy by analysing large volumes of patient data, often maintained on a monthly basis, which could be utilized to anticipate potential future diseases. Techniques such as Artificial Neural Networks (ANN), Random Forest, and Support Vector Machines (SVM) are widely applied to predict heart conditions. Diagnosing and predicting heart diseases remain a considerable challenge for both doctors and hospitals globally. To mitigate the high mortality rate associated with these diseases, efficient and rapid detection methods are essential. Machine learning and data mining techniques hold a crucial role in this context. Researchers are accelerating efforts to develop machine learning-based software that can assist doctors in both predicting and diagnosing heart diseases. This research project aims to leverage machine learning algorithms to predict the likelihood of heart disease in patients.

Index Terms- Neural Network, Machine Learning, Supervised Learning, Support Vector Machine, Random Forest.

I. INTRODUCTION

Machine learning involves the extraction of implicit and valuable information from data, a rapidly expanding field with increasing applications. It encompasses various classifiers in supervised, unsupervised, and ensemble learning categories, which can enhance accuracy and prediction capabilities for datasets. This knowledge can be applied to our Heart Disease Prediction System (HDPS), potentially benefiting many individuals. Cardiovascular diseases, a broad category encompassing numerous heart conditions, are increasingly common. The World Health Organization estimates that CVDs account for approximately 17.9 million deaths globally each year, making them the primary cause of adult mortality. This project aims to predict individuals at risk of heart disease by analysing their medical histories, identifying those with symptoms like chest pain or hypertension, thus aiding diagnosis with minimal testing and more effective treatment strategies.

The focus is on three data mining techniques: (1) Logistic Regression, (2) K-Nearest Neighbour (KNN), and (3) Random Forest Classifier. Our project achieved an accuracy of 87.5%, surpassing previous approaches that relied on a single technique. By employing multiple data mining techniques, we enhanced HDPS's accuracy and efficiency. Logistic regression, a supervised learning method, uses discrete values to classify whether a patient is likely to have cardiovascular disease based on medical attributes like age, gender, chest pain, and fasting blood sugar levels. Using a dataset from the UCI repository, which includes patients' medical histories, our model predicts heart disease likelihood based on 14 medical attributes. These attributes are processed using the three algorithms, with KNN achieving the highest accuracy at 88.52%. Ultimately, this cost-effective method helps classify patients at risk for heart disease.

II. RELATED WORK

In [1], researchers utilized data from the UCI repository to predict heart disease, applying KStar, J48, SMO, Bayes Net, and Multilayer Perception with WEKA software. Based on various performance factors, SMO achieved 89% accuracy, and Bayes Net achieved 87%, both outperforming KStar, Multilayer Perceptron, and J48 techniques using k-fold cross-validation. However, the accuracy results remained insufficient. In [2], data from Kaggle was used to predict stroke patients through Artificial Neural Networks (ANN) and Support Vector Machines (SVM), with accuracies of 81.82% for ANN and 80.38% for SVM on training data, and 85.9% and 84.26% for ANN and SVM, respectively, on the test dataset. In [3], researchers analysed data from the UCI repository, evaluating Naive Bayes, KNN, Decision Tree, and ANN algorithms, with ANN achieving the highest accuracy at 85.3%.

International Journal of Scientific Research & Engineering Trends
Volume 10, Issue 6, Nov-Dec-2024, ISSN (Online): 2395-566X

III. METHODOLOGY

The primary goal of this method is to predict the likelihood of heart disease, enabling early detection and timely intervention. Our approach utilizes a range of data mining techniques and machine learning algorithms, including Naïve Bayes, k-Nearest Neighbour (KNN), Decision Tree, Artificial Neural Network (ANN), and Random Forest, to make predictions based on various health parameters.

To conduct the analysis, we use Jupyter Notebook within Anaconda Navigator, an open-source platform that supports the implementation of multiple machine learning algorithms via library imports. Additional libraries can be downloaded using the Anaconda prompt. This tool enables us to write and execute live code, visualize data, perform data processing, and create graphs.

1. Dataset for Implementation

We selected a pre-existing dataset from the UCI Machine Learning Repository to predict heart disease. This dataset comprises 14 attributes, listed as follows:

- Age
- Sex
- Chest pain type (CP)
- Resting blood pressure (bps) in mm Hg
- Serum cholesterol level in mg/dL
- Fasting blood sugar (fbs) levels
- Resting electrocardiographic results
- Maximum heart rate achieved (thalach)
- Exercise-induced angina (exang)
- ST depression (oldpeak) induced by exercise relative to rest
- Slope of the peak exercise ST segment (slope)
- Number of major vessels colored by fluoroscopy (ca)
- Thalassemia (thal)
- Target variable (presence of disease: 0 or 1)

2. Data Splitting

The dataset is divided into training and testing subsets, with 75% used for training and 25% allocated for testing. We also applied data normalization techniques to manage and remove any NaN (Not a Number) values.

3. Visualization of Data

The performance and accuracy of each experiment are assessed using standard metrics such as true positive (TP) rate, true negative (TN) rate, precision, recall, and F-measure. These metrics are derived from the confusion matrix, often referred to as a predictive classification table, and provide a basis for comparing the effectiveness of the selected algorithms.

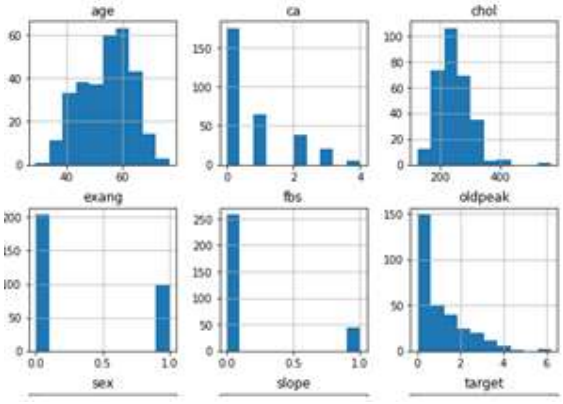

Figure 1: Machine Learning

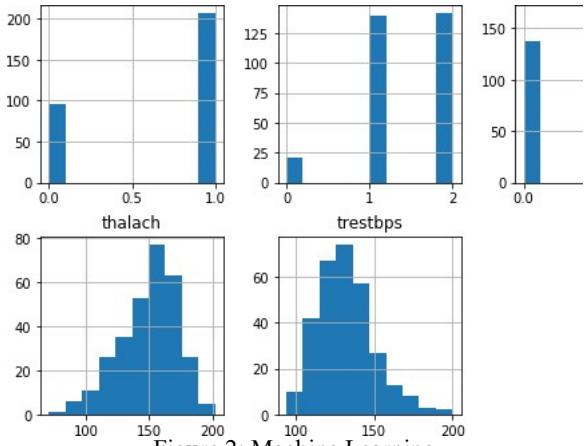

Figure 2: Machine Learning

Algorithms Used in Experiments

1. k-Nearest Neighbor (KNN)

KNN is a widely used machine learning algorithm, especially effective when working with continuous parameters. This algorithm classifies data points by predicting the "nearest neighbors" and is favored for its simplicity and speed. It can address both classification and regression tasks. In our case, KNN uses the heart disease dataset to classify whether an individual is at risk of heart disease. The algorithm calculates distances between data points on a graph to determine the proximity of points, thus helping classify individuals based on factors like age, gender, and other health parameters. KNN does not require a separate training phase for model generation, as the model is created directly from the dataset.

In visualizations, patients identified as having heart disease are marked by red dots, while those without are represented by blue dots.

Confusion Matrix

The accuracy of KNN largely depends on the chosen distance metric and the value of K. This accuracy is assessed through the confusion matrix, which provides insights into the

algorithm's precision. data is used in testing stage. It stores all the cases and then classifies new data according to the nearest neighbor.

KNN has two stages:

- Find the k number of instances in the dataset
- Use the k instances to find the nearest neighbor.

Visualization

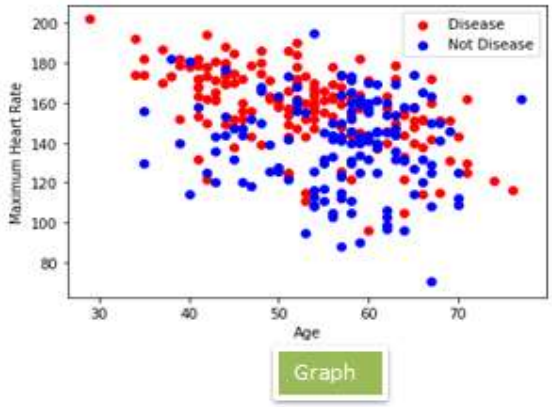

Figure 3: Disease Prediction Graph

The performance and accuracy of each experiment are assessed using standard metrics such as true positive (TP) rate, true negative (TN) rate, precision, recall, and F-measure. These metrics are derived from the confusion matrix, often referred to as a predictive classification table, and provide a basis for comparing the effectiveness of the selected algorithms

.

Algorithms Used in Experiments

1. k-Nearest Neighbour (KNN)

KNN is a widely used machine learning algorithm, especially effective when working with continuous parameters. This algorithm classifies data points by predicting the "nearest neighbors" and is favored for its simplicity and speed. It can address both classification and regression tasks. In our case, KNN uses the heart disease dataset to classify whether an individual is at risk of heart disease. The algorithm calculates distances between data points on a graph to determine the proximity of points, thus helping classify individuals based on factors like age, gender, and other health parameters. KNN does

KNN Model Result

	precision	recall	f1-score	support
0	0.90	0.79	0.84	33
1	0.85	0.93	0.89	43
accuracy			0.87	76
macro avg	0.87	0.86	0.86	76
weighted avg	0.87	0.87	0.87	76

not require a separate training phase for model generation, as the model is created directly from the dataset.

In visualizations, patients identified as having heart disease are marked by red dots, while those without are represented by blue dots.

Confusion Matrix

The accuracy of KNN largely depends on the chosen distance metric and the value of K. This accuracy is assessed through the confusion matrix, which provides insights into the algorithm's precision

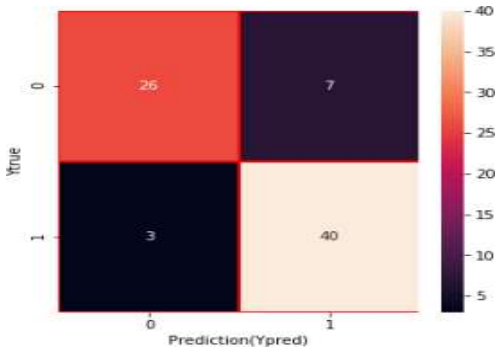

Figure 4: Confusion Matrix

It shows that it has 31 true negative rate, 4 false positive, 5 false negative while true positive cases are 36.

This model achieves an accuracy of 88%.

Decision Tree

A decision tree is a supervised learning classifier that is intuitive and easy to interpret. It can handle both numerical and categorical data. Structurally, a decision tree resembles a tree with internal nodes, branches, and leaf nodes. Each branch represents values within the dataset, internal nodes perform tests on specific attributes, and leaf nodes represent the predicted class or final outcome. The classification process follows a path from the root node down to a leaf node, determined by the predictive attributes and established rules.

Naive Bayes Model Result

	precision	recall	f1-score	support
0	0.86	0.89	0.87	35
1	0.90	0.88	0.89	41
accuracy			0.88	76
macro avg	0.88	0.88	0.88	76
weighted avg	0.88	0.88	0.88	76

It shows that it has 26 true negative rate, 7 false positive, 3 false negative while true positive cases are 40.

We achieved an accuracy of 87% by implementing this model.

Artificial Neural Network (ANN):

Artificial Neural Networks are computational models inspired by the human brain's neural structure. They consist of numerous nodes interconnected through directional links, where each node serves as a processing unit, and the connections represent causal relationships among them【4】.

ANNs are highly useful in clinical decision-making, assisting doctors in analysing data and making decisions with improved efficiency and accuracy.

A neural network begins with an input layer, where each input node connects to nodes in hidden layers, which may then link to an output layer. ANNs are designed to simulate and process information similarly to the human brain. With self-learning capabilities, they can continuously improve as more data is provided, resulting in progressively better outcomes. layer. This classification technique is becoming potent tool in data mining and may be utilized for different purposes in descriptive and predictive data mining [4]. The sample artificial neural network is shown in Figure below.

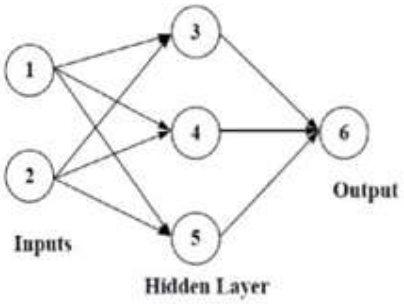

Figure 5: ANN

We achieved an accuracy of 87% by implementing this model.
Artificial Neural Network (ANN):

Artificial Neural Networks are computational models inspired by the human brain's neural structure. They consist of numerous nodes interconnected through directional links, where each node serves as a processing unit, and the connections represent causal relationships among them【4】. ANNs are highly useful in clinical decision-making, assisting doctors in analysing data and making decisions with improved efficiency and accuracy.

A neural network begins with an input layer, where each input node connects to nodes in hidden layers, which may then link to an output layer. ANNs are designed to simulate and process information similarly to the human brain. With self-learning capabilities, they can continuously improve as more data is provided, resulting in progressively better outcomes.
Confusion Matrix:

It shows that it has 30 true negative rate, 2 false positive, 8 false negative while true positive cases are 36.

This model gives accuracy of 87 %.

Naive Bayes
Naive Bayes is a classification algorithm grounded in Bayes' theorem. According to the Naive Bayes classifier, the presence of a specific characteristic in a class is assumed to be independent of other characteristics. This independence assumption makes it a strong classifier for predicting heart disease. Naive Bayes calculates the posterior probability of each class by using the conditional probability of the data. The formula is as follows:

P(C|X)=P(X|C)·P(C)P(X)P(C|X) = \frac{P(X|C) P(C)}{P(X)}P(C|X)=P(X)P(X|C)·P(C)

In this equation, XXX represents the instance to be predicted, and CCC is the class value. This formula helps determine the probability of the instance belonging to a particular class based on its features.

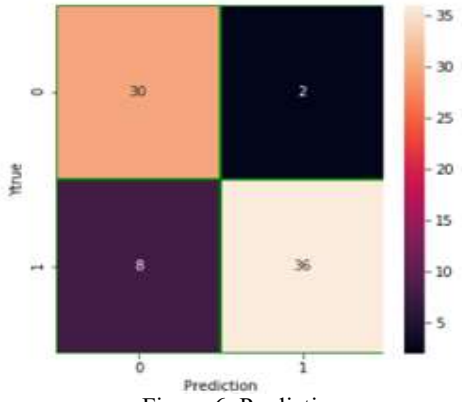

Figure 6: Prediction

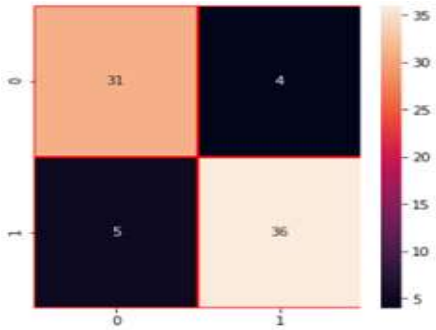

Figure 7: Confusion Matrix

Commonly used decision tree algorithms are CART, ID3, C4.5, J48 and CHAID are very important in the prediction of diseases [3].

Confusion Matrix

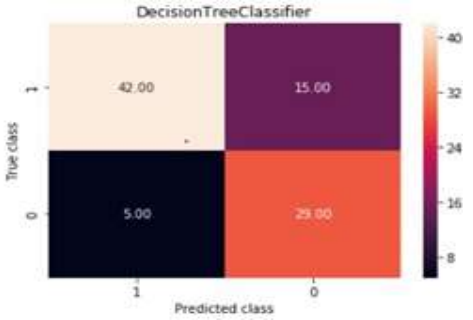

Figure 8: Confusion Matrix

It shows that it has 42 true negative rate, 15 false positive, 5 false negative while true positive cases are 29.

Decision Tree Result

	precision	recall	f1-score	support
0	0.85	0.66	0.74	44
1	0.74	0.89	0.81	47
accuracy			0.78	91
macro avg	0.79	0.78	0.78	91
weighted avg	0.79	0.78	0.78	91

This model gives accuracy of 78 %.

Random Forest

Random Forest is a supervised learning algorithm that can be used for both classification and regression tasks. Known for its flexibility and user-friendliness, Random Forest builds multiple decision trees using randomly selected data samples. Each tree produces a prediction, and the final result is determined by a majority vote among the trees. The more trees in the forest, the more robust the model. Our implementation of Random Forest achieved an accuracy of 82%, as shown in the classification report..

Confusion Matrix

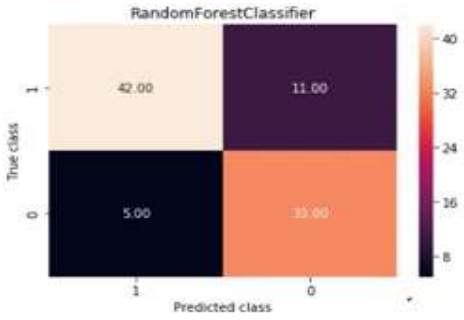

Figure 9: Confusion Matrix

Random Forest Result

	precision	recall	f1-score	support
0	0.87	0.75	0.80	44
1	0.79	0.89	0.84	47
accuracy			0.82	91
macro avg	0.83	0.82	0.82	91
weighted avg	0.83	0.82	0.82	91

We have achieved accuracy of 82 % as shown in classification report.

IV. DISCUSSION

Our study focuses on applying data mining techniques within healthcare, specifically for predicting heart disease. We conducted a series of experiments on a heart disease dataset, applying five distinct data mining algorithms to identify which provides the highest accuracy for predicting heart disease. The primary goal of these experiments was to evaluate and compare the effectiveness of KNN, Neural Networks, Decision Tree, Naive Bayes, and Random Forest for heart disease prediction.

Comparison of Implemented Algorithms

In our study, we used machine learning algorithms on heart disease patient data to evaluate each algorithm's predictive accuracy. Following the implementation, we compared the algorithms' performance using various metrics, including Accuracy, True Positive, False Positive, False Negative, True Negative, and the ROC Curve. The purpose of this comparison is to identify the algorithm with the highest accuracy and most reliable performance in predicting heart disease. The results of each algorithm are summarized in the table below.

Table 1.1: Comparing values.

Algorithms	Accuracy	TN	FP	FN	TP
KNN	0.87	26	7	3	40
ANN	0.87	30	2	8	36
Naïve Bayes	0.88	31	4	5	36
Decision Tree	0.78	42	15	5	29
Random Forest	0.82	42	11	5	33

Table 1.1 illustrates that the highest accuracy achieved on the dataset is 88%, while the lowest accuracy is 78%. The Naïve Bayes algorithm provides the highest accuracy, whereas the Decision Tree yields the lowest.

In terms of other performance metrics, the True Positive (TP) rates are as follows: KNN - 40, ANN - 36, Naïve Bayes - 36, Decision Tree - 29, and Random Forest - 33. This indicates that KNN has the highest TP rate, while Decision Tree has the lowest. Conversely, the Decision Tree shows the highest False Positive (FP) rate of 15, while ANN has the lowest FP rate of 2.

Based on this comparison, Naïve Bayes, KNN, and ANN perform well as they have similar accuracy rates and relatively high TP rates. KNN has an FP rate of 7, Naïve Bayes has an FP rate of 4, and ANN has the lowest FP rate of 2.

Given the critical nature of heart disease, which is responsible for millions of deaths worldwide, maintaining a high TP rate and a low FP rate is essential for accurate and timely diagnosis. Early and accurate detection can greatly improve treatment outcomes, making these metrics crucial in evaluating algorithm performance. Accuracy is also significant for identifying heart disease patients accurately.

V. CONCLUSION

Conclusion

Our study focused on using data mining techniques in healthcare, specifically for detecting heart disease, a life-threatening condition. We applied several data mining algorithms, including KNN, Neural Networks, Decision Tree, Naïve Bayes, and Random Forest, to a dataset of heart disease patients. Performance was assessed based on metrics such as Accuracy, True Negative (TN), False Positive (FP), False Negative (FN), and True Positive (TP) rates across the different algorithms.

We conducted five experiments using the same dataset to predict heart disease, with the results of all implemented algorithms summarized in a table for easier comparison. The experiments indicate that Naive Bayes achieved the highest accuracy at 88%, followed by ANN and KNN, both at 87%. Our findings suggest that data mining can be effectively applied within the healthcare industry to aid in early-stage prediction and diagnosis of diseases.

Future Work

Further research should focus on improving classification accuracy by incorporating advanced algorithms such as Bagging, Support Vector Machine, or Decision Table. Evaluating each algorithm's predictive performance and applying the proposed system to specific areas of interest could be beneficial. Adding additional features and adjusting algorithm parameters could enhance accuracy in future implementations.

In future studies, more comprehensive heart disease data and data reduction techniques could be utilized to improve prediction quality. To achieve better predictive results, high-quality datasets without inconsistencies should be used. This tool could serve as a dedicated resource for stakeholders to make informed decisions.

REFERENCES

1. Ujma Ansari, Jyoti Soni, Dipesh Sharma, Sunita Soni. "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", 258493784_Predictive_Data_Mining_for_Medical_Diagnosis_An_Overview_of_Heart_Disease_Prediction. March 2011Data Mining in Healthcare for Heart Diseases
2. C. Beyene, P. Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques", https://www.researchgatenet/publication/323277772_Survey_on_prediction_and_analysis_the_occurrence_of_heart_disease_using_data_mining_techniqus, 118(8):165-173 • January 2018
3. Muhammad Usama Riaz, SHAHID MEHMOOD AWAN, ABDUL GHAFFAR KHAN, "PREDICTION OF HEART DISEASE USING ARTIFICIAL NEURAL NETWORK", https://www.researchgate.net/publication/328630348_PEDICTION_OF_HEART_DISEASE_USING_ARTIFICIAL_N EURAL_NETWORK. October 2018
4. Umair Shafique, Irfan Ul Mustafa, Haseeb Qaiser, Fiaz Majeed, "Data Mining in Healthcare for Heart Diseases",https://www.researchgate.net/publication/274718934_Data_Mining_in_Healthcare_for_Heart_Diseases. March 2015.
5. Komal Kumar Napa, G.Sarika Sindhu, D.Krishna Prashanthi, A.Shaeen Sulthana, "Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifiers", https://www.researchgatenet/publication/340885231_Analysis_and_Prediction_of_Cardio_Vascular_Disease_using_Mac hine_Learning_Classifiers, April 2020.
6. Hossam Meshref, "Cardiovascular Disease Diagnosis: A Machine Learning Interpretation Approach",https://www.researchgatenet/publication/338428682_Cardiovascular_Disease_Diagnosis_A_Machine_Learni ng_Interpretation_Approach, January 2019.
7. Jabbar Akhil, Shirina Samreen, "Heart disease prediction system based on hidden naïve Bayes classifier", https://www.researchgatenet/publication/309735105_Heart_disease_prediction_system_based_on_hidden_naive_Bayes_c lassifier, October 2016.