Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Performance Comparison of Machine Learning Algorithms in Heart Disease Prediction with Enhanced Accuracy through Hyper parameter Tuning

Nisreen Ryadh Hamza[a,*], Farah Jawad Al-Ghanim[b]

[a]*Computer Science Department, College of Computer Science and Information Technology, University of Al –Qadisiyah Al-Diwaniah, Iraq.*

E-mail: nesreen.readh@qu.edu.iq

[b]*Computer Science Department, College of Computer Science and Information Technology, University of Al –Qadisiyah Al-Diwaniah, Iraq.*

E-mail :farah.jawad@qu.edu.iq

ARTICLE INFO

Article history:
Received: 08 /04/2025
Rrevised form: 29 /04/2025
Accepted : 07 /05/2025
Available online: 30 /06/2025

Keywords:

Heart Disease,

Machine Learning,

Random Forest,

Decision Tree,

K-Nearest Neighbors ,

Support Vector Machine,

Naive Bayes.

ABSTRACT

Heart disease, which frequently results from blockage of the coronary arteries, the blood channels that supply the heart with oxygen-rich blood, is still one of the top causes of death globally. Plaque and fatty deposits that accumulate along the arterial walls are the main causes of this blockage, which makes the arteries narrow and limits blood flow. Even with cardiac disorders' seriousness and potentially fatal consequences, early detection remains a significant challenge in the medical field, often due to the complex and subtle nature of early symptoms. This diagnostic difficulty highlights the need for advanced computational tools that can support clinical decision-making. In this context, this study investigates the application of machine learning algorithms to heart disease risk prediction. Five artificial intelligence models—Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), Decision Tree (DT), and Naive Bayes (NB)—were tested on a dataset of 1,888 records and 14 attributes. In order to increase data quality and guarantee that models can learn patterns efficiently, preprocessing was used to clean and prepare the data first. This was followed by hyper parameter optimization for the SVM and KNN models. The aim of hyper parameter optimization is to maximize the model's performance on the data. With an accuracy of 96.30%, Random Forest outperformed the other models under evaluation. Decision Tree came in second with 95.59% , SVM with 94.55% ,KNN with 0.9425 and Naive Bayes with 0.6966 . These studies demonstrate how machine learning may be used to identify cardiac illness early on by identifying intricate patterns in data and providing more precise results than conventional techniques.

MSC..

https://doi.org/10.29304/jqcsm.2025.17.22180

1. Introduction

When the coronary arteries, which are blood channels that provide the heart with oxygen and blood, get clogged, a condition called heart disease develops. Fatty debris and plaque build-up inside the coronary arteries are the causes of coronary heart disease [1]. Despite their lethal nature, cardiac illnesses are difficult to detect in their early stages. Researchers are constantly searching for better detection techniques that might swiftly identify cardiac

∗Corresponding author: *Nisreen Ryadh Hamza*

Email addresses: *nesreen.readh@qu.edu.iq*

Communicated by ‘sub etitor’

disease in a person because the existing approaches are not as effective as one would like in diagnosing the disease in its early phases due to efficiency and computational time.[2]. Additionally, it can be quite challenging to spot any heart disease symptoms until a person starts to have chest pain or complains of breathing problems. It might be quite challenging to treat the illness and save the life of the individual by that point. It is particularly true when professional medical experts and cutting-edge technology are not easily accessible. It will be crucial to develop a tool that can identify such a life-threatening condition early on in order to help doctors identify it in its early stages and to spare an individual from experiencing this agony. According to the Centers for Disease Control and Prevention (CDC)[3], heart attacks can be hard to diagnose until a person experiences symptoms including chest discomfort, neck or upper back pain, indigestion, nausea, vomiting, etc.; heart failure symptoms like exhaustion, foot swelling, etc.; or an arrhythmia like palpitations. Smoking, high blood pressure, and high cholesterol are the three main risk factors for this deadly disease, and over fifty percent of US have at least a single of these conditions. About 697,000 Americans lost their lives to heart illness in 2020, according to statistics given for the year. That indicates that one in five Americans lost their lives to heart disease in 2020! Cardiovascular problems can also be caused by a poor diet, being overweight or obese, having diabetes, drinking too much alcohol, etc. In 2020, heart disease, commonly referred to as cardiovascular sickness, was the primary cause of mortality in the US, with a staggering 928,741 recorded, according to the American Heart Association's article, Heart illness and Stroke Statistics – 2023 Update. In the United States,41.2% of all fatalities were attributable to coronary artery disease.. Stroke came in second with 17.3% of all cardiovascular disease-related deaths, high blood pressure claimed 12.9% of lives, heart failure claimed 9.2%, and some disease that spread in arteries claimed 2.6% of lives. According to data collected in 2020, the age-adjusted mortality rate from cardiovascular diseases in the US was 224.4 per 100,000. Turning to the financial side of cardiovascular disease, the data showed that, in 2018 and 2019, a total of $251.4 billion was spent directly to combat the disease, and that, A staggering $407.3 billion was spent over the course of the two years to combat cardiovascular disease [4]. Two of the key forces behind the development of technology in the medical and healthcare industries are data science and machine learning. Most companies in this industry are progressively adjusting and incorporating data science and machine learning into their systems to increase their chances of finding patterns for different ailments. Medical businesses can identify trends, draw conclusions from data, and take on a variety of computationally challenging jobs by implementing machine learning techniques, according to the Neptune Blog[5]. The correlation between several characteristics and traits of patients with the designated illness can be ascertained by data scientists using machine learning tools. Consequently, this enables physicians to comprehend the disease's trends and develop more effective preventative treatments for their patients. It is crucial to take into account the risks associated with handling a sizable collection of data .An organization's credibility and reputation may suffer if millions of patients' data are stolen through system hacking. It is essential that developers build a model using a language that has facilities that offer protection over any external risks in order to prevent such situations from emerging. Python is one of the top five languages used to construct healthcare systems, according to a 2017 Stack Overflow Developer survey, the results of which are displayed in the figure in Belitsoft's article "Python in Healthcare"[6]. The study aims to evaluate the performance of a set of machine learning models in predicting the risk of heart disease using real medical data, in order to identify the most accurate and effective model for potential use in medical applications.The structure of the study is as follows: Section 2 provides an overview of the relevant literature. Section 3 describes the methodology. Section 4 presents and discusses the findings. Finally, Section 5 concludes the paper and suggests directions for future research.

2 .Literature Review

According to research, There are several methods for using artificial intelligence (ML) to forecast heart illness. According to their study, Machine Learning-Based Heart Disease Prediction System, The Artificial Neural Network algorithm is one of the suggested methods to forecast cardiac illness, according to Ranjit Shrestha and Jyotir Moy Chatterjee. This is primarily because it can lower the cost of the diagnosis by allowing for the use of an alternative test to decide whether to diagnose HD (Heart Disease)[7].

Santosh Kumar , Samir Patel, and Devansh Shah presented a model that was based on learning with supervision . It utilizes the Cleveland database, a UCI repository of individuals suffering from heart disease. Estimating the patients' risk of developing heart disease is the aim of this investigation. KNN produces the highest accuracy score, according to the results[8].

The paper presents an approach to Enhance Human Cardiac Disease Prediction by Hindawi Abdul Saboor and his colleagues claimed that doctors mostly employed the auscultation approach to distinguish between normal and abnormal heart sounds using machine learning algorithms. . Despite giving doctors the ability to identify a variety of heart disease types, the auscultation method did not always accurately classify and clarify the various sounds because that depends on the doctors' knowledge and practices, which can only be acquired through extensive

examinations. A dataset with 303 records and 76 attributes was used to train and assess the model using the cross-validation tenfold technique[2].

A new approach and technique were proposed by ZAHER AL AGHBARI, AMAL AL ALI, MARIAM ELJAMIL, AND AHMED M. KHEDR to manage the mining of dispersed medical data sources at various locations utilizing Rules of Association. In the suggested approach, the agent either travels to each location and completes local tasks or exchanges a few simple summaries with other agents to compute the global association rule. The suggested strategy accomplishes the same goals as if the data were sent and combined at a single location while protecting the privacy of patient data[9].

The Multi-Layer Per and the KNN are two Machine Learning "ML" techniques that are described in a published article by Madhumita Pal along with her associates . Among the 13 important characteristics in the dataset are gender, age, chest discomfort, Resting blood pressure ,cholesterol level , and so on (1 indicates cardiovascular disease and 0 indicates no cardiovascular disease)[10].

Sushanth Ganjala, Nikhil Debbarma, and Jyoti Kiran used a range of artificial intelligence approaches to identify the best reliable and accurate strategy for heart disease diagnosis. A lot of the stress associated with determining the classifier's probability of accurately diagnosing heart disease is reduced by the model that is offered. While boosting medical treatment, it reduces costs [11].

Avinash Sharma, B. K. Agrawal, and Deepika Arora focused on examining a variety of methods for predicting cardiac diseases. The most successful algorithms, according to the results, are Random Forest and XGBoost[12].

A number of difficulties still exist in spite of the wide range of earlier research that used machine learning algorithms to diagnose cardiac disease. These include the use of sparse or tiny datasets, the absence of thorough method comparisons, and the inadequate examination of many performance indicators. Furthermore, a lot of research ignored pragmatic issues like data privacy and actual implementation difficulties. The potential for direct application in actual healthcare settings is further limited by the lack of clinical interpretation for model results. By thoroughly comparing five distinct machine learning algorithms using a trustworthy dataset, this study tackles a number of issues raised by earlier research. The performance of some models was improved, demonstrating a focus on obtaining accurate and trustworthy results. The study offers a helpful standard for future research and emphasizes the great potential of machine learning techniques in assisting with the early diagnosis of heart disease. These efforts are considered the main contributions of the research, providing valuable insights and benchmarks for further advancements in the field Furthermore, the database was most recently updated in November 2024, suggesting that it is current and that no prior studies have been conducted using it.

3. Methodology

We will explain the methodology followed in this research with a set of steps explained in detail in the fig1:

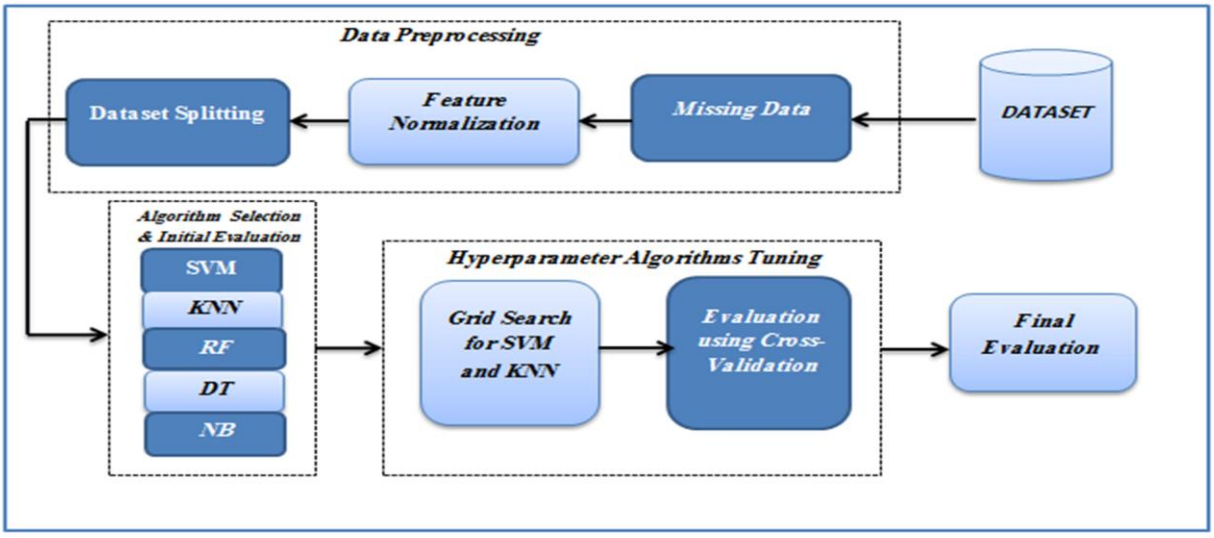

Fig. 1. Steps of the Methodology

Explanation of the steps of the methodology in detail:

3.1. Data Collection

With 1,888 records, this collection combines five publicly accessible heart disease datasets. Combining these datasets offers a stronger basis for developing machine learning algorithms that forecast the likelihood of a heart attack. The dataset's diversity, well-balanced classes, and robust model performance suggest that it is adequate for accurate heart disease prediction, even with its size. There are 14 features in this dataset that are known to increase the risk of a heart attack. Even without genetic or medical history information, the 14 features are sufficient for accurate prediction, as they are clinically validated and strongly correlated with heart disease. Since these datasets are available to the general public, they are open-source and not sourced from a private healthcare facility. Developing neural network (machine learning models) that aim to prevent and identify cardiac disease early is a wonderful fit for it. To guarantee data integrity, missing data has been eliminated from the records. Numerous machine learning algorithms, including classification models like Decision Trees "DT" and Neural Networks "NNs", can be used with this dataset [13]. Below is a detailed description of each feature (show in fig.2):

- Age: Age of the patient (ex.: 50)
- Sex: Gender of the patient, 1 denotes a man, and 0 a woman.
- CP: Type of chest discomfort. Atypical "1", non-anginal pain "2", and asymptomatic " 3".
- Resting Blood Pressure is represented as trestbps.
- Chol: The numerical value of the serum cholesterol level .
- FBS: Greater than "120 mg/dl" . Truth "1", whereas Falsity "0".
- Restecg: Normal "0", abnormal "1", and left ventricular hypertrophy"2".
- Halach: reached the highest possible heartbeat..
- Exang: Exercise-induced hypertension. Affirmative "1", and No"0".
- Old peak: Depressive symptoms of ST caused by exercise as opposed to rest.
- Slope: The decline of the segment with ST at the highest exercise level. Up sloping "0", flat "1", and down sloping "2".
- CA: The diversity of significant arteries with fluoroscope colours (0–3).
- Types of thalassemia: One represents a normal defect, two a fixed flaw, and three a reversible defect.
- target: The outcome variable (risk of a heart attack). Values: 1 indicates a higher risk of a heart attack, whereas 0 indicates a lower risk.

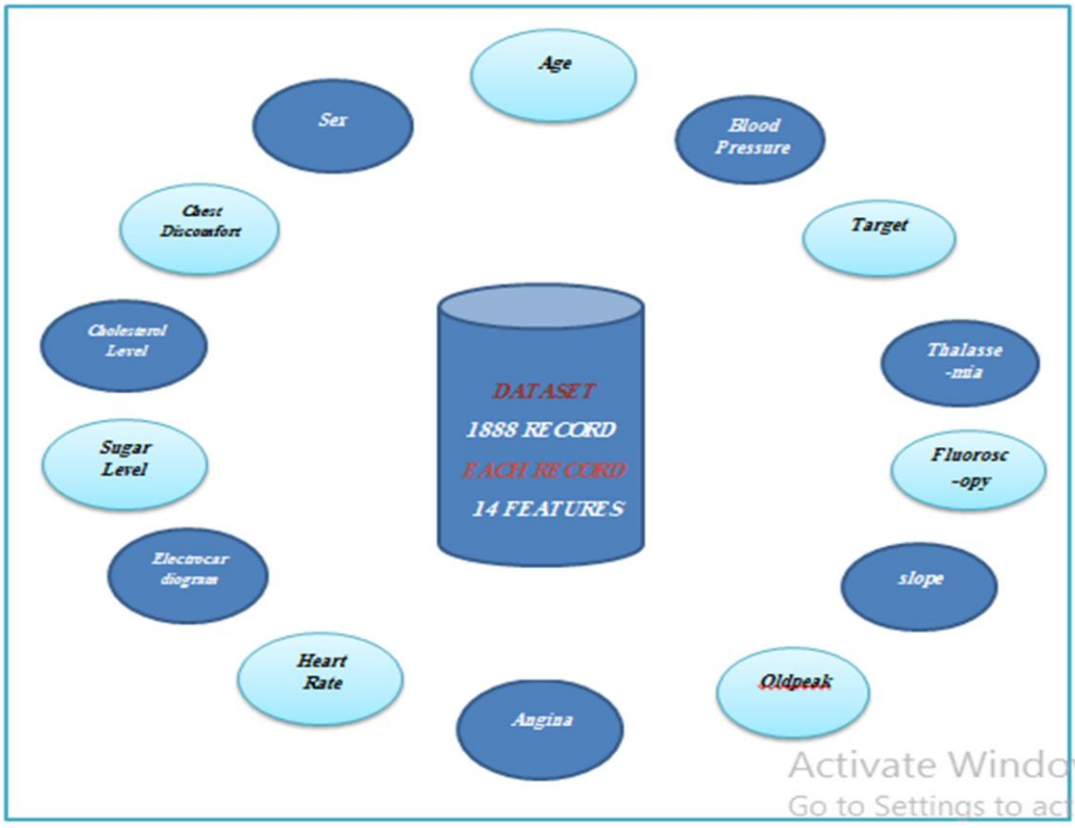

Fig.2. Features in each record in data base

3.2. Data Pre-processing

The following processes were used to guarantee data quality and get it ready for machine learning algorithms:

- Handling Missing Data: Records that included incorrect or missing values (e.g., "?") were deleted. This guarantees that each feature is clear and useful.
- Feature Normalization: For feature-scale-sensitive techniques, like SVM and KNN, the normalization (standardization) of the features is an essential step.
- Dataset Splitting: 70% of the data collection utilized for training, and 30% utilized for testing.

3.3. Artificial Intelligence Techniques (Machine Learning)

For evaluation, five machine learning techniques were used. Every algorithm was trained and evaluated separately.

a. SVM stands for Support Vector Machines.

In a M-dimensional plane where "M" is the characteristics number , SVM or SVC aims to construct a hyper plane to accurately categorize each point of data. SVM can describe intricate decision boundaries and works well with high-dimensional data[16].fig.3 shows the best hyper plane graphs for Support Vector Classifiers

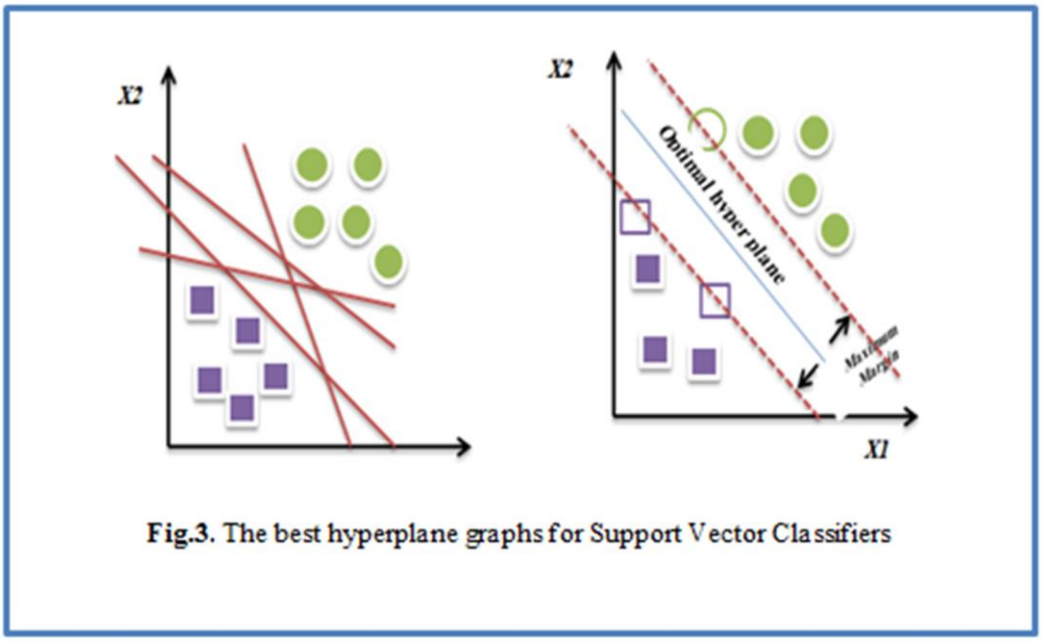

Fig.3. The best hyperplane graphs for Support Vector Classifiers

b. KNN stands for K-Nearest Neighbours

The Nearest Neighbours Classifier classifies, or predicts, how the individual data points will cluster based on the K - nearest neighbours' closeness to a specific point. Due to it retains the information and then conducts a task on it throughout the procedure for categorization rather than learning from a training set right away , this technique is sometimes referred to as the lazy-learner algorithm.[15] The method will separate the newly acquired dataset into discrete groups that are comparable to the freshly added data to the database. The dataset is only saved while the algorithm is in the training phase. Fig.4 below shows an illustration of the Nearest Neighbours Classifier for reference.

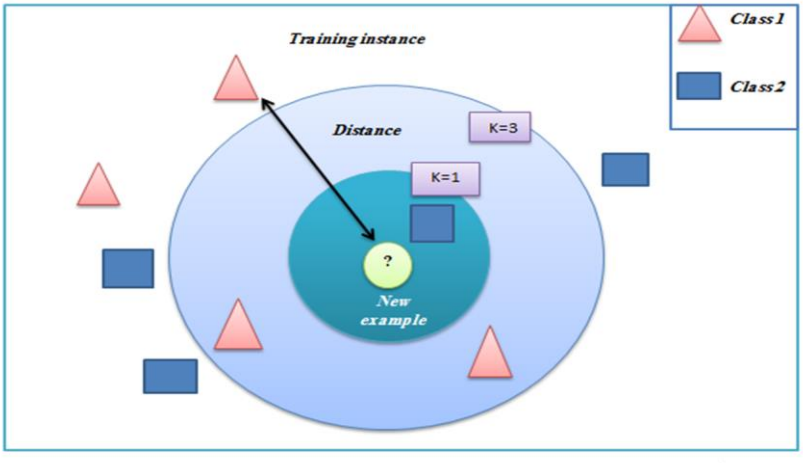

Fig. 4. The Nearest Neighbors Classifier.

c. RF stands for Random Forest

The Random Forest Classifier "RF" will construct many decision trees and consider various aspects from the overall features in order to enhance estimates for the complete data set. Every tree in this procedure will forecast a category, and the category that retains the most forecasts will be the model's prediction. Fig.5 serves as a reference to illustrate this idea[17]:

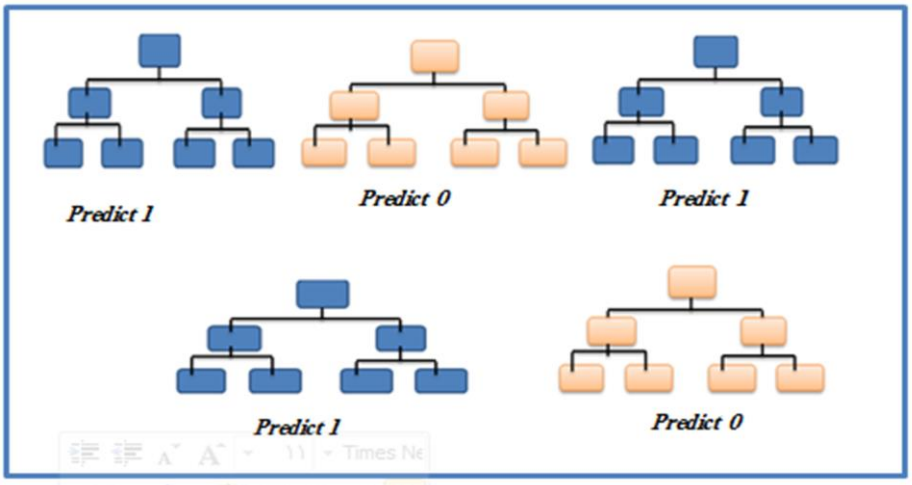

Fig.5: Random Forest Classifier using Decision Trees / *Predict 1*

d. DT stands for Decision Tree

In order to develop a predictive framework that may be applied to draw inferences from a variety of observations, the Decision Tree Classifier will employ a classification tree. The findings that are obtained based on several feature-based splits are displayed using a tree-like structure. The Decision Tree Classifier process starts at the root node, as seen in Fig.6 below. This root node is then split up into other nodes, which are decision nodes. Ultimately, we get to the conclusion, or leaf node, and make a prediction based on the choice that is made. Decision Trees are interpretable and computationally efficient [14].

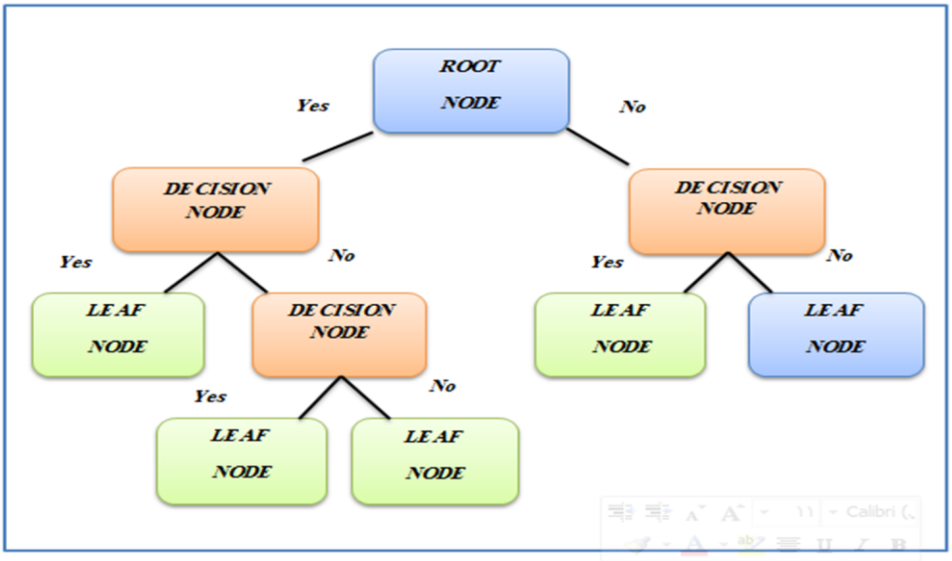

Fig.6. Decision Tree

e. NB stands for Naive Bayes

A straightforward probabilistic classifier, the Naive Bayes approach generates a set of probability by counting the "frequency" & "value" combinations in a set of data. Considering the class variable's value, the algorithm makes the assumption that all qualities are independent and applies the Bayes theorem. Although the technique is characterized as naive ,because this independent condition assumption is rarely true in practical applications, it often performs well and learns fast across a range of supervising categorization issues. The foundation of the Naive Bayesian classifier is the notion of total probability and the Bayes theorem (show in fig.7).[18].

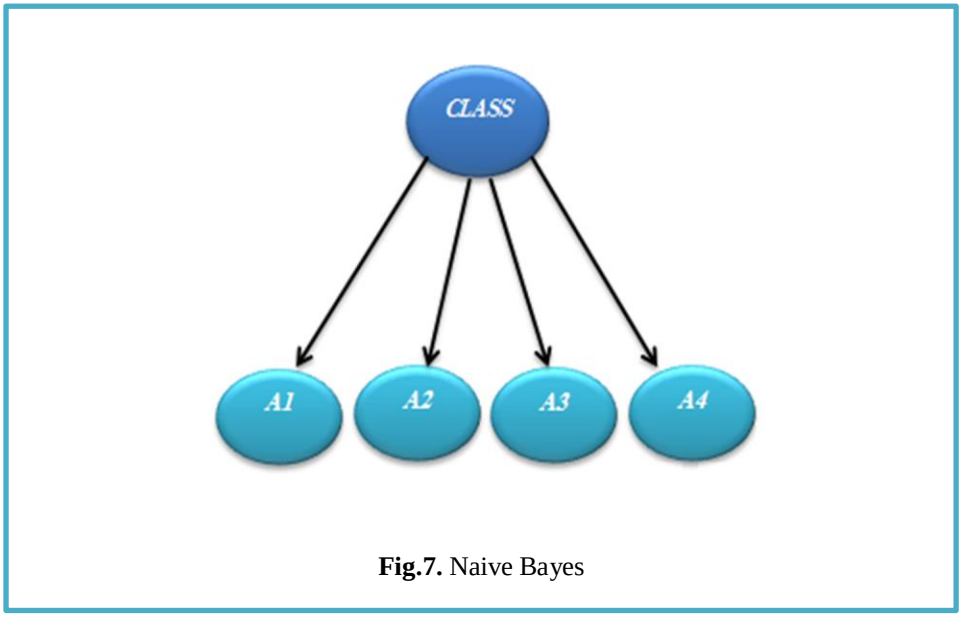

Fig.7. Naive Bayes

3.4. Hyperparameter Algorithms Tuning

In order to optimize configuration and improve model performance, the following hyperparameter tuning strategies were used:

- Grid Search for SVM and KNN: C and gamma for the RBF kernel were optimized using SVM Grid search. The optimal k (number of neighbors) and weighting technique were found using KNN Grid search.
- Evaluation using verification by cross-checking : The use of five-fold cross- validation was employed during the tuning of hyper parameters.to evaluate each model. This guarantees unbiased and reliable parameter selection.

3.5 Final Evaluation

Four evaluation metrics—Accuracy, Precision, Recall, and F1-Score—were used to gauge the machine learning models' performance. The percentage of all forecasts that were correctly classified is known as accuracy. Precision measures the model's ability to prevent false positives by dividing the number of accurately predicted positive observations by the total number of anticipated positives. Recall measures the sensitivity of the model by showing the percentage of real positive examples that were accurately detected. The F1-Score is a balanced metric that takes into consideration both false positives and false negative.

4. Results and Discussion

Based on an actual dataset, this section presents the findings from the models used to forecast heart disease. The objective is to examine each model's accuracy and efficacy, emphasizing the variations in the algorithms used and the variables affecting their results. Based on the results, the discussion also sheds light on each model's fitness for the prediction task. The precision, recall, and F1-score of machine learning algorithms for two classes are compared in the table 1. SVM and KNN perform the worst overall, followed by Random Forest and Decision Tree. Particularly in Class 0 recall and Class 1 precision, Naive Bayes has the lowest performance. For this task, RF and DT are the most dependable options overall.

Table 1 - Training and validation losses and F1 scores over epoch.

Algorithm	PRECISION "CLASS 0"	RECALL "CLASS 0"	F1-SCORE "CLASS 0"	PRECISION "CLASS 1"	RECALL "CLASS 1"	F1-SCORE "CLASS 1"
Support Vector Machine "SVM"	0.96	0.93	0.94	0.93	0.97	0.95
K-Nearest Neighbors "KNN"	0.98	0.89	0.93	0.91	0.98	0.94
Random Forest "RF"	0.98	0.95	0.96	0.95	0.98	0.96
Decision Tree "DT"	0.97	0.93	0.95	0.94	0.98	0.96
Naive Bayes	0.73	0.59	0.65	0.68	0.79	0.73

The results demonstrate that "RF" attained the maximum level of accuracy (96.30%), followed by "DT" (95.59%) and SVM (94.55%). KNN also performed well (94.25%), while "NB" lagged behind (69.66%). The results presented emphasize how crucial it is to choose the right algorithms and adjust their hyper parameters. Random Forest's ensemble approach likely contributed to its superior performance. A summary of each model's accuracy rates is given in Table 2 below, which further demonstrates these conclusions

Table 2- Accuracy Comparison

Algorithm	Accuracy
SVM	0.9455
KNN	0.9425
RF	0.9630
DT	0.9559
NB	0.6966

The fig.8 uses confusion matrices to compare five models. SVM and Decision Tree are the next most accurate algorithms, after Random Forest. Errors were higher for KNN and Naive Bayes, whereas Naive Bayes performed the worst.

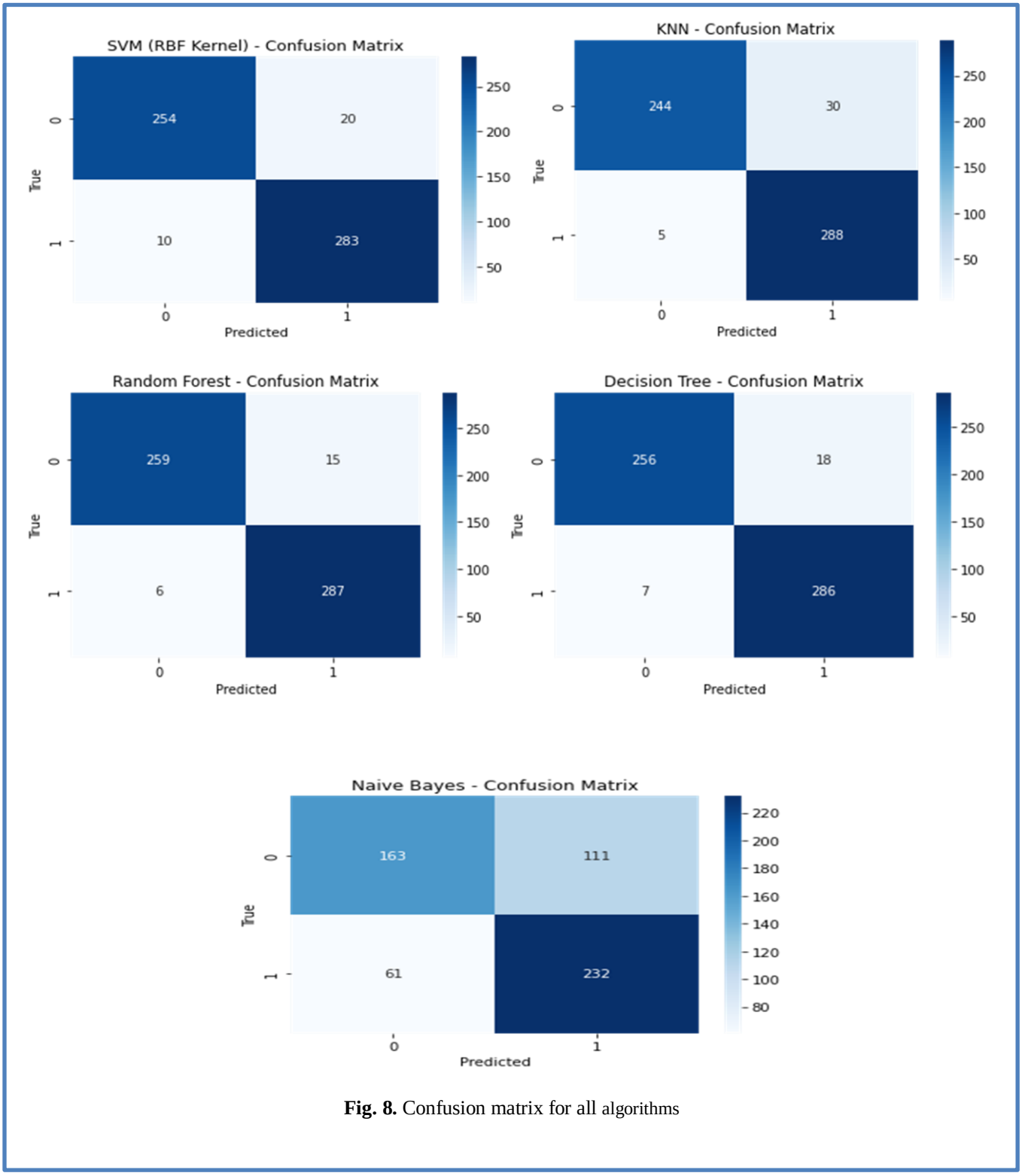

Fig. 8. Confusion matrix for all algorithms

The accuracy of several algorithms is readily compared in the chart (fig.9). The best accuracy was attained by Random Forest (96.30%),

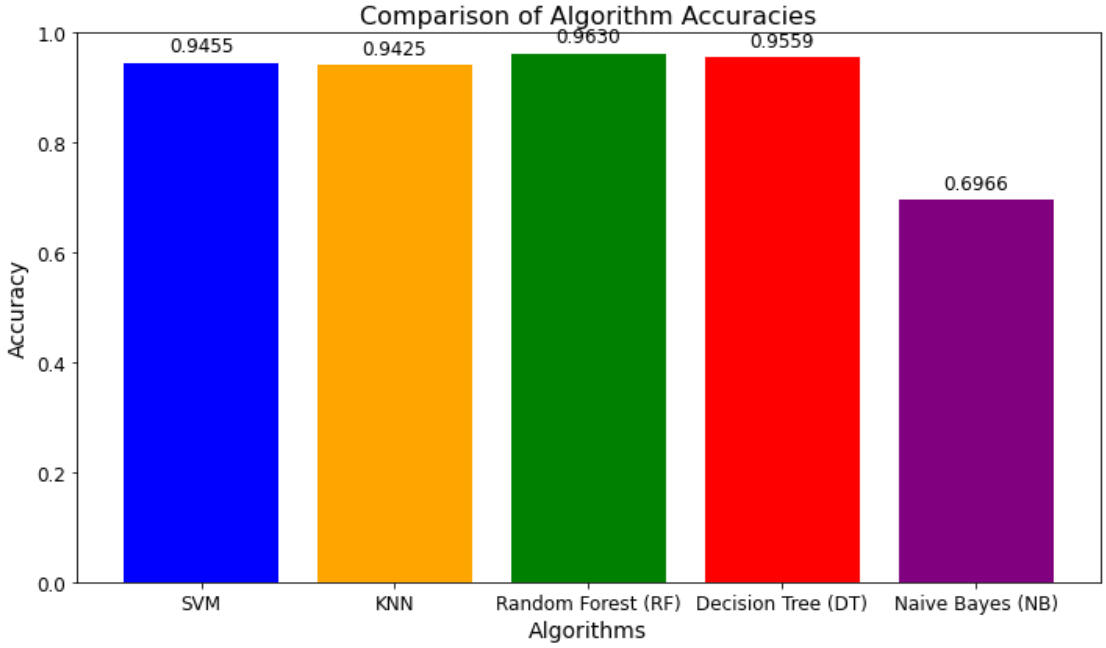

Fig. 9. Accuracy comparison charts

5. Conclusion and Future Works

Outperforming models like Decision Tree and SVM, this study showed that ensemble learning techniques, in particular Random Forest, produced the highest accuracy (96.30%) in predicting the risk of heart disease. Several models' performance was further improved by hyperparameter tuning, demonstrating the importance of algorithm improvement. These findings demonstrate the great potential of machine learning, particularly ensemble methods, to aid in precise and timely cardiac diagnosis. The dataset's diversity, balanced class distribution, and robust model performance suggest that it is enough for accurate heart disease prediction despite its small size. Additionally, because the datasets are publicly available and open-source, they provide a useful and repeatable basis for creating and evaluating machine learning models. No previous studies appear to have focused on or utilized the dataset employed in this research. Nonetheless, The lack of real-time clinical validation and the potential for data bias, however, are limitations of the study that should be addressed in future research by adding real clinical data, diversifying datasets, and testing models in actual medical settings. Additionally, integrating explainable AI and advanced deep or hybrid models may enhance interpretability and clinical trust, laying the groundwork for wider adoption in healthcare.

References

[1] National Cancer Institute. (n.d.). NCI dictionary of cancer terms. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/coronary-heart-disease

[2] Saboor, A., Khan, M. A., Alghamdi, N. S., & Almotiri, S. H. (2022). A method for improving prediction of human heart disease using machine learning algorithms. Mobile Information Systems, Hindawi. https://doi.org/10.1155/2022/4721354

[3] Centers for Disease Control and Prevention. (2022, July 12). About heart disease. https://www.cdc.gov/heartdisease/about.htm

[4] American Heart Association. (2023). Heart disease and stroke statistics - 2023 update. https://professional.heart.org/en/science-news/heart-disease-and-stroke-statistics-2023-update

[5] Barla, N. (2022, July 21). Data science and machine learning in the medical industry. Neptune.ai. https://neptune.ai/blog/data-science-machine-learning-in-healthcare

[6] Shestel, A. (2021, March 18). Python in healthcare. Belitsoft. https://belitsoft.com/python/python-in-healthcare

[7] Shrestha, R., & Chatterjee, J. M. (2019). Heart disease prediction system using machine learning. LBEF Research Journal of Science, Technology and Management, 1(2).

[8] Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. International Journal of Computer Applications, 1, Article 345.
[9] Khedr, A. M., Al Aghbari, Z., Al Ali, A., & Eljamil, M. (2021). An efficient association rule mining from distributed medical databases for predicting heart diseases. IEEE. https://doi.org/10.1109/ACCESS.2021.3055203
[10] Pal, M., Parija, S., Panda, G., Dhama, K., & Mohapatra, R. K. (2022). Risk prediction of cardiovascular disease using machine learning classifiers. Open Medicine (Warsaw), 17(1). https://doi.org/10.1515/med-2022-XXXX
[11] Kiran, J., Debbarma, N., & Ganjala, S. (2023, April 26). Heart disease prediction using machine learning. Conference Paper.
[12] Arora, D., Sharma, A., & Agrawal, B. K. (2024, April 30). Assessing the impact of various machine learning algorithms for heart disease prediction. Conference Paper.
[13] Nazirkhan, F., & Rajiah, S. (2024). Heart disease prediction dataset. Kaggle. https://www.kaggle.com/datasets/mfarhaannazirkhan/heart-dataset/data
[14] SmartDraw. (n.d.). Decision tree – Learn everything about decision trees. https://www.smartdraw.com/decision-tree
[15] Webtunix Solutions. (n.d.). K-Nearest Neighbors Classifier. https://www.ris-ai.com/k-nearest-neighbors-classification
[16] Gandhi, R. (2018, July 5). Support vector machine - Introduction to machine learning algorithms. Medium – Towards Data Science. https://towardsdatascience.com/support-vectormachine-introduction-to-machine-learning-algorithms-934a444fca47
[17] Yiu, T. (2021, September 29). Understanding Random Forest. Medium – Towards Data Science. https://towardsdatascience.com/understanding-random-forest-58381e0602d2
[18] Dimitoglou, G., Adams, J. A., & Jim, C. M. (n.d.). Comparison of the C4.5 and a Naïve Bayes classifier for the prediction of lung cancer survivability.