Journal of Information Systems Engineering and Management
2025, 10(10s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

An Analytical Approach for Early Heart Disease Detection and Prevention Leveraging Machine Learning Techniques

*Md Imran Alam[1], Mohiuddin Ali khan[1], Huda Fatima[2], Haneef Khan[1], Sarfaraz Ahmed[1], Aasif Aftab[2], Mohammad Rafeek Khan[1], Shams Tabrez Siddiqui[2]

Department of Electrical and Electronics Engineering, College of Engineering & Computer Science, Jazan University, Saudi Arabia[1]
,Department of Computer Science, College of Engineering & Computer Science, Jazan University, Saudi Arabia[2]
**mimran@jazanu.edu.sa, makhan@jazanu.edu.sa, hsaadullah@jazanu.edu.sa, haneeskhan@jazanu.edu.sa, swahaj@jazanu.edu.sa, aaftab@jazanu.edu.sa,mokhan@jazanu.edu.sa,stabrez@jazanu.edu.sa*
Corresponding author : Md Imran Alam, mimran@jazanu.edu.sa

ARTICLE INFO

Received: 02 Nov 2024

Revised: 22 Dec 2024

Accepted: 04 Jan 2025

ABSTRACT

A global health issue caused by heart failure is long-term morbidity in most patients with the condition. These conditions should be identified as early as possible for a correct diagnosis, especially to enable proper therapeutic approach. This research aims to assess whether ML methods can develop an accurate first-stage diagnostic model in the detection of HF. On the basis of a set of clinical, demographic, and diagnostic data, various supervised learning methods were tested to predict heart failure for each patient. The mathematical foundation of this study includes the use of supervised learning models, each trained to approximate a mapping function f: X->Y, which are symbolized by the vectors X, to the outcomes Y, where Y = {0, 1}, where 0/1 is the absence/presence of heart failure. The diagnostic ability of models will determine their performance according to precision, an accuracy rate, recall, F1 score metric, and AUC-ROC. The main aim of this present study is therefore two folds; To establish how early the cardiac complications can be detected using different machine learning techniques so that preventive action can be taken. It aims at filling this gap through an evaluation of the literature to establish measures of handling imbalanced datasets. The accuracy of heart disease predictions improved with the implementation of seven machine learning techniques: Random forest, gradient boosting, logistic regression, decision tree, support vector machine (SVM), k-nearest neighbors (KNN), and Naive Bayes. In this work, a focus is given to the mathematical and computational aspects of machine learning methods in cases of early diagnosis and intervention of cardiac issues. It affords a sound support within the clinical settings for proactive and accurate healthcare involvement through the operationalization of predictive analytics into clinical decision making.

Keywords: Machine learning methods, diagnosis, metrics, Accuracy, Precision, Random forest, gradient boosting, logistic regression, Naive Bayes,datasets.

INTRODUCTION

Worldwide heart disease and other CVDs still rank as the number one killers with profound effects for health and medicine. Optimizing patient outcomes and reducing the disease's prevalence necessarily involve early diagnosis. But there are issues with the efficacy, cost, and expandability of the current diagnostic frameworks [3]. Machine Learning has already had a substantial impact in medical and life sciences research. Diabetes is a metabolic condition characterized by persistently elevated levels of blood sugar that do not respond to insulin. Diabetic patients who are diagnosed early are more likely to live a healthier lifestyle [10]. The problems listed above have been addressed by recent developments in ML which presents new modalities in diagnosing health care. Being a subset of AI, machine learning uses algorithms to search through large information and arrive at diagnosing and predicting and those and more often better than conventional statistical models to improve diagnosis and treatment, healthcare providers are utilizing wearable technology that is built on the IoT. An abundance of Internet-enabled sensors and devices have recently emerged. The Internet of Health Things, has numerous applications in the healthcare sector [21]. The healthcare system may relieve some of the burden caused by cardiovascular disease if early detection and prevention efforts are encouraged. There has been significant progress in improving the effectiveness and prediction of heart disease predictions using cloud computing (CC) and machine learning techniques in recent years [8]. Recent advancements in the transformer class of models and DL have enhanced the potential of ML systems to diagnose cardiac illness. By using proposed vision based transformer models, the ability to diagnose from these ECG images is much improved [5]. This shows that DL methods hold the potential to work with large amount of unstructured medical data. Other scalable models for healthcare cloud are also considered and Prediction models have also been introduced in cloud allowing fast data processing and real time analysis through a powerful cloud based infrastructure for cardiovascular disease prediction. Such innovations effectively demonstrate that new technology can assist in unraveling barriers that have for many years hindered the provision of healthcare services. The benefits of using ML for the prediction of cardiac illness are improved accuracy of diagnosis, shorter survey time, and improved decision-making [9]. There are still issues such as data quality, no association with clinical operations, and largely unclear algorithms. To address these matters, both data scientists and physicians should collaborate, and create AI that is comprehensible [7]. Because of demographic factors, diet, and increase exposure to physical hazards, heart diseases are emerging and more lethal and is a major global health issue. Valvular heart disease one of the categories of cardiovascular disorders is increasing globally, particularly in countries with lower and middle income [1]. The global population is aging and rheumatic heart disease and other unidentified diseases are worsening the situation. Over the course of time, new advancements have been made in the medical treatment of congenital heart disorders thus altering its epidemiology of its presentation in adult patients' prediction and living standards [6]. Structurally, heart disease and mental health are intertwined creating complications on how doctors may diagnose and treat them. Highlighted the point that anxiety, depression, and stress as key mental disorders are contributors to CHD in both hypothesized directions [2]. For this reason, it is important that any illnesses are treated and approaches to curbing illnesses need to consider the effects on the body and the mind of the patient. As per the latest statistics available in Heart Association of America, cardiovascular strokes still hold the rank of leading cause of death around the world [11]. They illustrate why it's so important to look for new methods of diagnosing diseases and avoiding them. Most cardiovascular diseases produce no symptoms till they are well advanced and hence early diagnosis is vital in reducing mortality and morbidity. One of the promising and inspiring technologies that can be applied to identify the development of cardiovascular pathologies in its early stages as well as predict its future, is machine learning (ML). Explored and discussed the exploration of ML techniques emphasizing their capacity to assess complex data, to consider persons who are at risk, and to improve the diagnostic adenomatous polyposis [4]. Another benefit of adopting the technology is that new intervention strategies and tailored approaches can be developed by applying the available massive data for identifying patterns which are not distinguishable to stochastics. Beginning with the increasing worldwide concern for cardiovascular disease and its link to mental health, this overview aims to build knowledge of the revolutionary potential of machine learning in improving early diagnosis and treatment.

This research outlines the benefits and drawbacks of utilizing machine learning for the diagnosis of cardiac ailments. After that, we looked at seven predictive models based on machine learning to enhance the detection of cardio-vascular and vascular diseases: Some of the stated methods include the Decision Tree, Gradient Boosting, Support Vector Machine and K-Nearest Neighbors, Naive Bayes and Logistic Regression. Hence based on the methodologies used for pre-processing of datasets for machine learning research, current study involves dimensionality reduction, converting categorical variable into numerical, normalization and scaling strategies to handle missing values and balancing of datasets. Lastly, we contrasted and assessed the effectiveness of various techniques for cardiovascular disease diagnosis of healthcare-related cases using ML.

LITERATURE REVIEW

Let's see how machine learning is changing the face of healthcare: Especially in cardiovascular diseases, the effective early detection can be offered with the help of new approaches proposed by the ML. To offer some epistemological input on the use of ML approaches for the prediction and diagnosis of heart disease, therefore, this review captures the current results. Another major challenge in cardiovascular diseases datasets is the problem of class imbalance. This takes place when the number of healthy instances outdo the number of ill cases, this causes a disposition to bias the outcomes. Current methods amounting to minimizing class disparity were discussed, and new methods such as resampling methods and cost-sensitive learning were proposed [18]. Their study also highlighted that different ensemble methods that were used together, including SMOTE with gradient boosting techniques provided higher sensitivity and specificity to the tests. These approaches ensure that good model performances are achieved, which is of great importance in clinical areas, where the cost of false negatives is high. Cardiovascular diagnosing has been enhanced because ML algorithms enable precise quick diagnosis by using the predictive modeling. Emphasized high accuracy of different types of ML methods to analyze demographic and clinical data, such as random forest, decision tree and support vector machines [14]. Another focus was made on ensemble learning techniques that use characteristics of a number of algorithms to produce more accurate predictions. Likewise, we considered the role of DL models to ML in the function of identifying cardiovascular illness [12]. They also found that RNNs, and CNNs, provided a better accuracy in diagnostic when fed with complex data such as ECG and image data. To successfully apply these models in the clinical environment, one needs to conceptualize such models with help of specific explainable AI (XAI) frameworks. Even today, the prediction of cardiovascular diseases requires the use of the conventional ML method. In particular, demonstrated the use of logistic regression, Naive Bayes and k-Nearest Neighbors or k-NN algorithms for the prediction of risk indicators of heart diseases [19]. However, feature selection was particularly important for enhancing the effectiveness of these models when working with small datasets, according to their study. Moreover, investigated the actual world-based clinical data with ML models comprising decision trees and SVMs [17]. They also established that by doing comparison, ensemble methods like the random forest were better than the individual methods. Used CNNs for evaluating ECG and image data to diagnose cardiac diseases since they considered CNNs as a method for determining the presence of disease [13]. For example, their work reveal that recall and precision can be significantly higher with CNN-based models for identifying the existence of coronary artery disease. Scalable solutions for CV diagnosis have been achieved through integrating ML with cloud-based platforms but the authors highlighted concerns with data quality and model explain-ability as critical to clinical implementation. The rapid forecasting of heart failure health status and subsequent execution of appropriate measures to address this worldwide concern necessitate an effective machine learning technique. The primary treatment for heart failure is medication, however exercise is increasingly being recognized as a successful supplement therapy [15]. This method enables simple integration of predictive systems in off-site healthcare facilities as well as enhance access. The suggested machine learning models should reduce the time for diagnosis and carrying out of tests and surveys [16]. They provided beneficial findings that can be potentially improved to faster and more accurate cardiac disease prediction systems through the implementation of state-of-art, low complexity algorithms. However, a several challenges remain when attempting to predict cardiac disease using ML; there is still a lot of work to be done. Enlisting into panel of clinical procedures, model interpretability of the predictions, and data quality issues are part of this [14]. To overcome these challenges and achieve enhanced comprehensiveness in healthcare, novel trends including federated learning and explainable AI form helpful solutions.

DATASET AND METHODOLOGY

To pursue the above laid out study objectives, the following approaches are applied as follows: Those enable people to know more about the different aspects of sudden cardiac arrest; therefore, models for diagnosing and receiving such problems become more effective. Figure 1 provides a comprehensive summary of the research methodology employed in this study.

DATASET DESCRIPTION

The dataset utilized in this investigation has been obtained from Kaggle [20]. Fourteen attributes constitute this dataset. Table 1 provides an illustration of each characteristic detail. Records from 1025 patients, including 713 males and 312 females of various years of age have been reviewed in the dataset. Out of these, normal patients accounted for 499 (48.68%), and cardiac disease patients accounted for 526 (51.32%). We also find that out of these patients, 300 are males and 226 are female having cardiac disease.

SN	**Attribute Name**	**Description**
1	age	Age is measured in years
2	sex	Male = 1/Female = 0
3	cp	Chest pain type having (4 values)

4	trestbps	Blood pressure in resting is measured (in mm Hg on admission to the hospital)
5	chol	Serum cholesterol is measured in mg/dl
6	fbs	Fasting blood sugar should be > 120 mg/dl (1 = true & 0 = false)
7	restecg	Electrocardiographic results in resting
8	thalach	Achieved the maximum heart rate
9	exang	Exercise-induced angina (1 = yes & 0 = no)
10	oldpeak	ST depression induced by exercise relative to rest
11	slope	The slope of the peak exercise ST segment
12	ca	Number of major vessels (0–3) colored by fluoroscopy
13	thal	1 = normal; 2 = fixed defect; 3 = reversible defect
14	Target (Class)	0 = no disease and 1 = disease

Table 1: Key Attributes, Explanations and Value Ranges

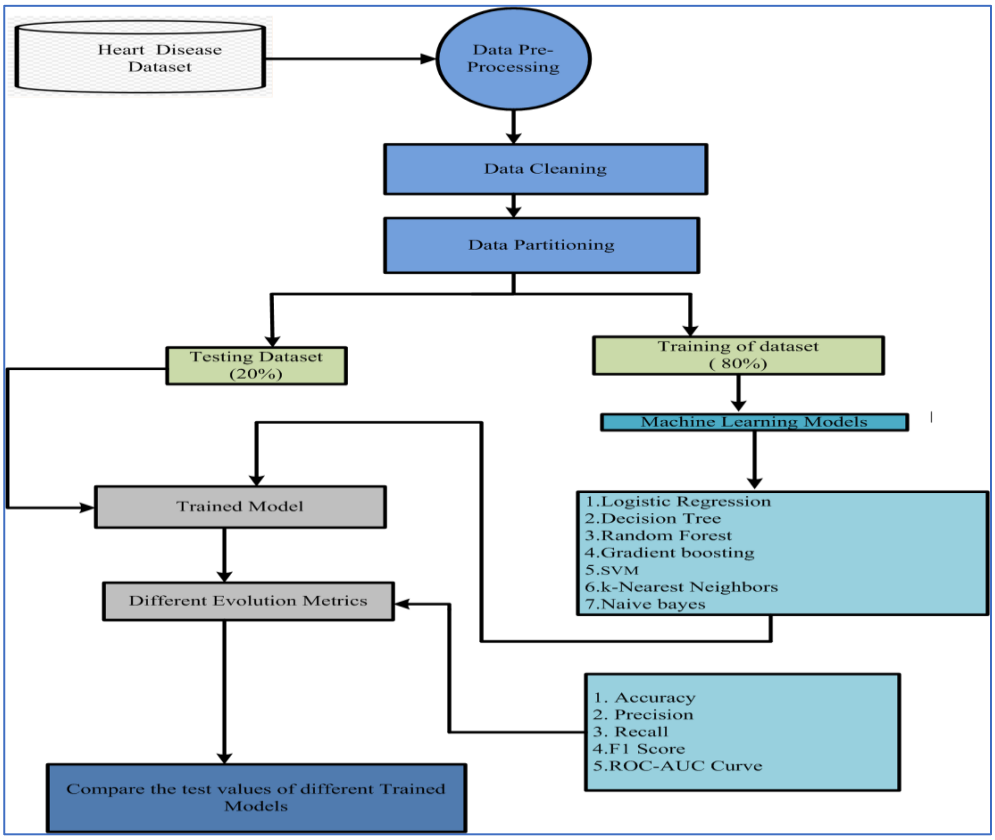

Figure 1 Proposed Model for performance analysis of Different Machine Learning Techniques

DATASET PRE-PROCESSING

The below preparation procedures were carried out to guarantee the datasets were appropriate for machine learning:

i. Handling Missing Values: K-NNearest Neighbors were applied when using model-based imputation or mean/median imputation were used to address missing data.

ii. Balancing the Dataset: To handle internal class imbalance for the classes where it appropriate, Synthetic Minority Over-sampling Technique (SMOTE) was used.

iii. Normalization and Scaling: Both continuous characteristics were either normalized into having zero mean and unit variance, or normalized with a range of 0-1.

iv. Encoding Categorical Variables: Other characteristics such as male or female and the type of pain were represented via one hot encoding.

v. Dimensionality Reduction: If redundancy is to be removed and the model has to be, more compact, techniques like PCA were used.

DATA PARTITIONING

Each of the datasets is broken down into its appropriate training, validation, and test sets, as shown in the following breakdown:

Training was conducted on 80% of the data set. In this study, hyperparameters tuning and the assessment of the training performance were done with 20% of the validation set data.

MACHINE LEARNING ALGORITHMS

This section includes seven prominent machine learning techniques: Random forest, logistic regression, SVM, decision tree, gradient boosting, KNN, and naive bayes are used in model development.

i. Logistic Regression: This algorithm deals with binary classification problem. It measures the probability that an input x is of particular class Encouraging the utilization of this probability measure makes the model more predictable.

Mathematical Model:

$$h_x(\theta) = \sigma(w^T x + b), \tag{1}$$

where:

- $\sigma(z) = 1 / (1 + e^{-z})$ it is sigmoid function.

- w: Weight vector.

- b: Bias term.

- x: Input feature vector.

Decision Rule: Predicted class: The estimated value of $h_x(\theta)$ is assigned as: 1 if $h_x(\theta) > 0.5$, otherwise 0.

Loss Function:

Binary cross-entropy loss:

$$L(\theta) = -(1/m)\, \Sigma\, [y\log(h_x(\theta)) + (1- y) \log(1- h_x(\theta)]. \tag{2}$$

ii.Decision Tree: It is a kind of learning algorithm which does not involve parameters and can be used in both classification and regression. It involves partitioning of the data into subsets for the sake of feature values, creating a structure that resembles a tree.

Splitting Metrics:

- Gini Index: $G = 1 - \Sigma\, p_i^2$, (3)

where p_i represents probability of class i.

- Entropy: $H = -\Sigma\, p_i \log_2(p_i)$. (4)

- Information gain(IG)= H(parent) – W (Average(H(children))).

iii. Random Forest: It is an ensemble learning procedure that builds large numbers of decision trees and uses them all in conjunction.

Model:

- Classification: These are the computational methods used to make the classification prediction; the mode is used to arrive at the predicted class.

- Regression: Predicted value is the mean of the tree outputs.

A bootstrap sample and feature subset selection are used to generate each tree.

iv.Gradient Boosting: To minimize the errors of the prior models, this boosting is an amalgamating method of machine learning that builds the prediction model by adding instants of weak learners' chiefly decision trees. It is often applied when working with the loss function when building regression as well as classification models by applying the gradient descent technique.

Suppose {(x_i, y_i)}_{i=1}^n be the dataset, where x_i stands for the input features and y_i for the goal.

i. Setting up: Start with a constant value ie the mean of the regression target: F_0(x) = the constant that minimizes Σ L(y_i, c).

ii. Updates Iteratively: where for m (number of iterations) = 1, 2 ... M.

Determine the residuals, or the loss function's negative gradient, as follows:

m_i^(m) = - [dL(y_i, F(x_i))/dF(x_i)] (5)

when evaluated at F(x) = F_(m-1)(x).

Fit the residuals to a weak learner (such as a decision tree) h_m(x): h_m(x) = r_i^(m). (6)

In order to update the model, add the weak learner's scaled predictions: The function Upgrade F_m is equal to function F_(m-1) plus the product of the learning rate, which is denoted by ν and the function h_m of the upgrade.

iii. Final Forecast: For regression, η is equal to F_M(x) which means that Bass has accounted for math in making his forecasts.

Perform categorization using an appropriate transformation of probability such as softmax.

Here the symbol F_m(x) is used in the paper to refer to the entire model constructed during m th iteration of the process. The basis function weak learner introduced at iteration m is regarded as h_m(x).ν controls the output of each weak learner and is known as learning rate.

Mean squared error in regression and log loss in classification, denoted as L(y, F(x)), are two such examples.

v. Support Vector Machine (SVM): This technique is an attempt to find best hyper-plane to split the data into different classes.

Model:

Decision boundary: $f(x) = w^T x + b$ (7)

Optimization:

In the case of least squares, we just try to minimize $(1/2)||w||^2$ and try to satisfy $yi(w^T xi \cdot b) \geq 1$ for all i

For soft-margin SVM:

$$\min ||w||^2/2 + C\, \Sigma\, \xi_i ||w|| \quad (8)$$

where $y_i(w^T x_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0$.

vi.K-Nearest Neighbors (KNN): The labels of a point's closest k neighbors are used to classify it in this lazy learning strategy.

Distance Metric:

Euclidean Distance: $d(x, x') = \sqrt{\Sigma(x_i - x'_i)^2}$ (9)

For classification purposes, the majority of class ‘K' nearest neighbors is taken into consideration.

vii.Naive Bayes: It is a probabilistic classifier which is based on statistical probability with using of Bayes' formula, and presuming features' independence from each other.

Model: $P(y|x) = (P(x|y)P(y)) / P(x)$ (10)

Assuming feature independence:

$P(x|y) = \Pi P(x_i|y)$. (11)

Predicted class:

$\hat{y} = \text{argmax\_y } P(y) \Pi P(x_i|y)$. (12)

PERFORMANCE METRICS FOR MODEL EVALUATION

The assessment of models plays a significant role in machine learning because it determines how close to reality forecasts made by properly trained models are. This leads to choices about implementation and improvement, and guarantees that models can successfully adapt to new information. The following performance Metrics for model evaluation was used in order to conduct an assessment on this study.

Here's the meaning of each symbol:

TP (True Positive): It finds the number of cases which were positively classified by the model, in other words, the number of times a person was described as having a disease when the fact is true.

TN (True Negative): It is the count of times the model was correct to predict no disease = true negative cases.

FP (False Positive): It represents the number of cases where the model got it wrong – that is, cases that were incorrectly classified to the positive class, for instance, in a disease prediction, the model labelled the person as having the disease while in real sense he or she does not which is also referred to as a “Type I alpha error.”

FN (False Negative): It represents the number of times the model said no disease where there actually was disease present in the patient. Type II error is another name for it.

i. Accuracy: It specifies the proportion of accurately identified occurrences, both positive and negative, relative to the total instances. It is an essential metric employed for the purpose of data classification.

Formula:

$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$ (13)

ii. Precision (Positive Predictive Value): It is the ratio of the correctly identified positive cases among all cases identified as positive. It shows the extent of positivity among all the positive predictions that has been made.

Formula:

$\text{Precision} = TP / (TP + FP)$ (14)

iii. Recall (Sensitivity, True Positive Rate): It shows the rate of truly positive instances out of all instances tagged as positive. It shows the degree at which the model captures the positive instances.

Formula:

$\text{Recall} = TP / (TP + FN)$ (15)

iv. F1-Score: It is the harmonic mean between precision and recall offering a unitary measurement of both. It is advantageous when one of the classes is dominated by the other depending on the different examples as follows;

Formula:

$$\text{F1-Score} = 2 * [\ (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})\ ]$$

Substituting the formulas for precision and recall:

$$\text{F1-Score} = 2TP / (2TP+FP+FN) \qquad (16)$$

v. ROC AUC (Receiver Operating Characteristic Curve, Area Under the Curve): The ROC AUC has counterparts in various fields since it estimates the area under the ROC curve. While discussing model performance, it is worth noting that a higher AUC value indicates better results.

Formula for AUC:

$$AUC = \int TPR(t)\, d(FPR(t)) \qquad (17)$$

Where:

Recall $TPR = TP / (TP + FN)$ or True Positive Rate

$FPR = FP / (FP + TN)$ Proportion of effectively negative cases in the population False Positive Rate

EXPERIMENTAL RESULTS

In the outcomes section of our study, we aim to demonstrate how each component of our innovative framework improves the precision of heart disease prediction. This analysis is vital to a comprehensive understanding of each aspect's contribution to achieving the researched objectives.

Analysis Based on performance metrics for each model:

i. Logistic Regression Performance Metrics

Performance Metrics	Values
Accuracy	0.85
Precision	0.84
Recall	0.86
F1-Score	0.85
ROC AUC	0.87

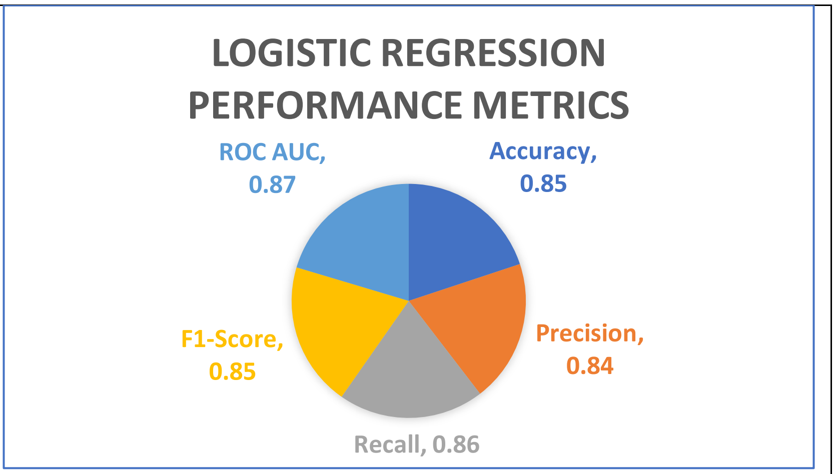

Figure 2: Logistic Regression performance metrics

Analysis: From figure 2, it can be analysed that Logistic Regression has a good score of F1 having values 0.85 & ROC AUC having values 0.87 means a better performance is observed from Logistic Regression. Hence, it shows that because Recall is slightly above precision, the algorithm can identify the true positives while having false positives.

ii. Decision Tree performance metrics

Performance Metrics	Values
Accuracy	0.79
Precision	0.78
Recall	0.81
F1-Score	0.79
ROC AUC	0.75

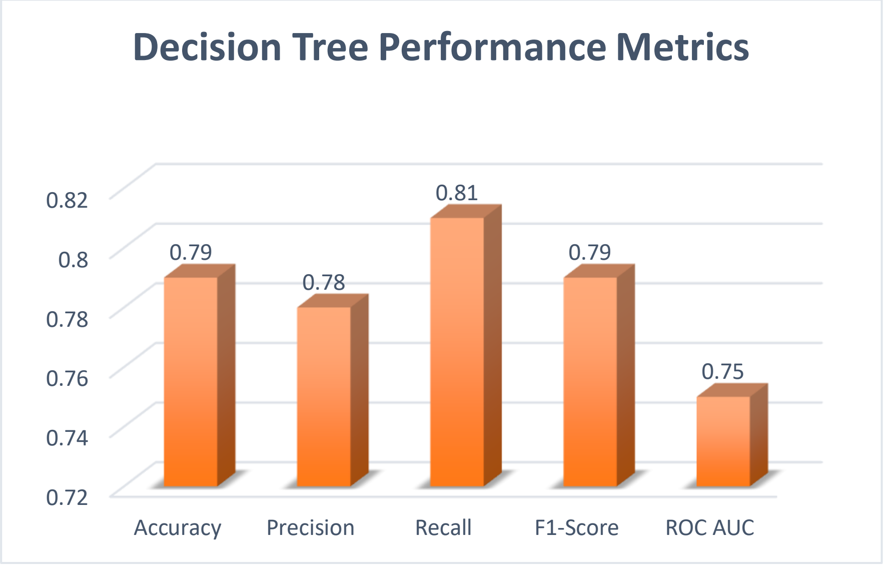

Figure 3: Decision Tree performance metrics

Analysis: It is observed from figure 3 that the Decision Tree has least merit than other models possessing ROC AUC of 0.75 confirms its poor discriminative power. Recall is high at 0.81 while precision is low, at 0.78 meaning while the algorithm may not miss many relevant tweets, it also flags some irrelevant ones. This may indicate that the model gets highly trained in the training data but it performs poorly once it meet's new, unseen data.

iii. Random Forest performance metrics

Performance Metrics	Values
Accuracy	0.88
Precision	0.87
Recall	0.89
F1-Score	0.88
ROC AUC	0.9

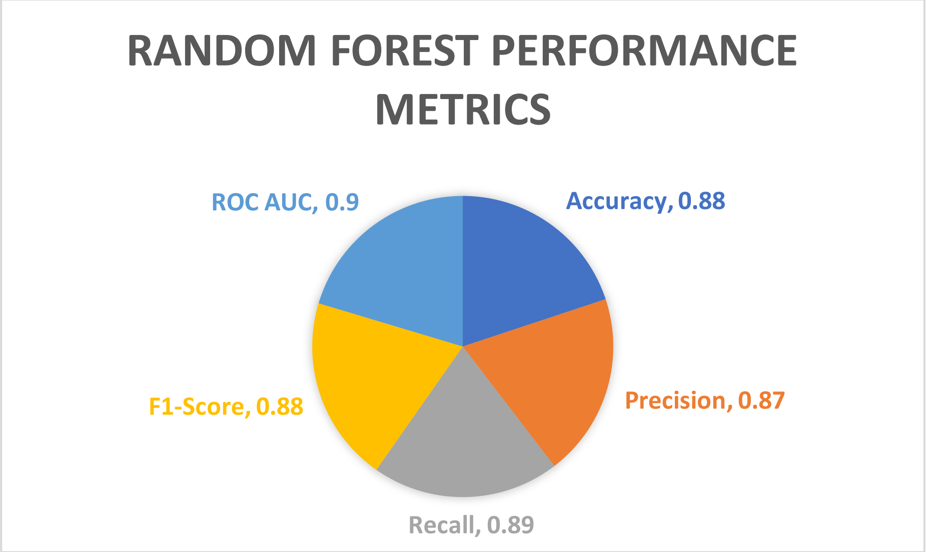

Figure 4: Random Forest performance metrics

Analysis: From figure 4 Random Forest is analyzed to have good balanced performance with all the metrics used. ROC AUC=0.90 clearly show excellent power of discriminative ability which are very useful for classification purpose. It has reasonably high precision, 0.87, and the recall rate of 0.89, which reduces the values of both false positives as well as false negatives.

iv. Gradient Boosting Performance Metrics

Performance Metrics	Values
Accuracy	0.95
Precision	0.93
Recall	0.92
F1-Score	0.92
ROC AUC	0.97

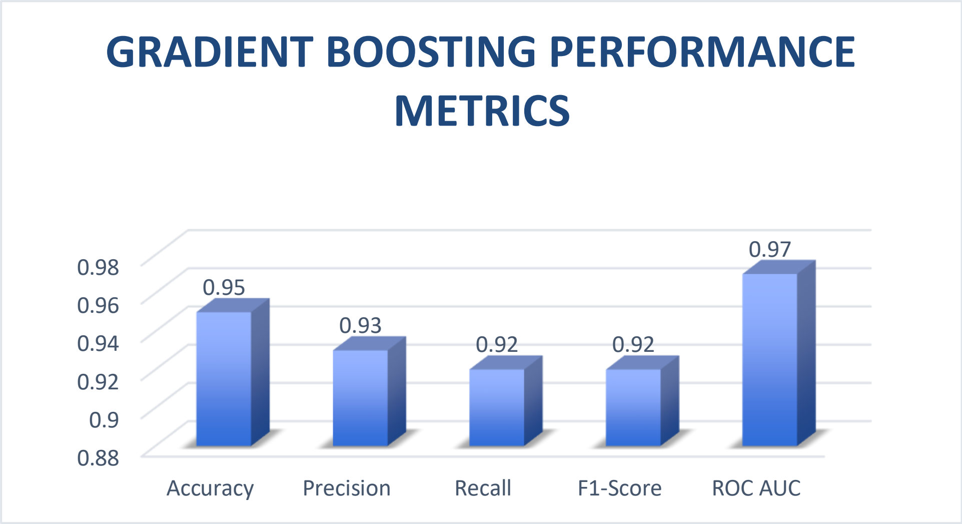

Figure 5: Gradient Boosting Performance Metrics

Analysis: From figure 5 it is seen that the model classifies 95% of the instances right hence proving that the model is most reliable. A ROC AUC of 0.97 confirms the strong performance of the proposed model in terms of class discrimination. Thus, achieving a good trade-off where false positives are balanced in numbers with false negatives, there is a desire for utilization of precision and recall with equal measures of balance. Therefore, the Gradient Boosting Algorithm yields a superior accuracy for heart disease prediction with robust precision and recall that are applicable in clinical decision support.

v. Support Vector Machine performance metrics

Performance Metrics	Values
Accuracy	0.86
Precision	0.85
Recall	0.86
F1-Score	0.85
ROC AUC	0.88

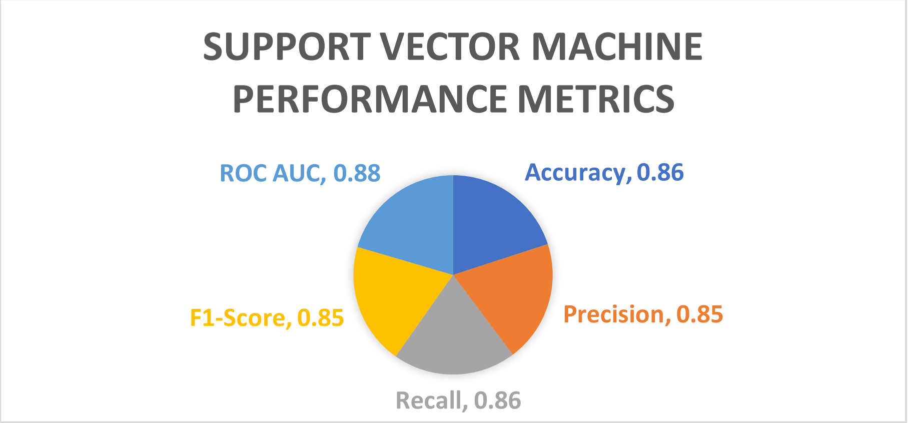

Figure 6: Performance Metrics of Support Vector Machine

Analysis: In figure 6, similarly, it is discussed that, the outcome of SVM is also significantly efficient with balanced metric as compared to Logistic Regression. It it has high ROC AUC of 0.88 which is indicative of high capability of discriminating between classes. It is preferably less sensitive to the problem of over fitting compared to the conventional models.

vi. K-Nearest Neighbors performance metrics

Performance Metrics	Values
Accuracy	0.82
Precision	0.81
Recall	0.83
F1-Score	0.82
ROC AUC	0.8

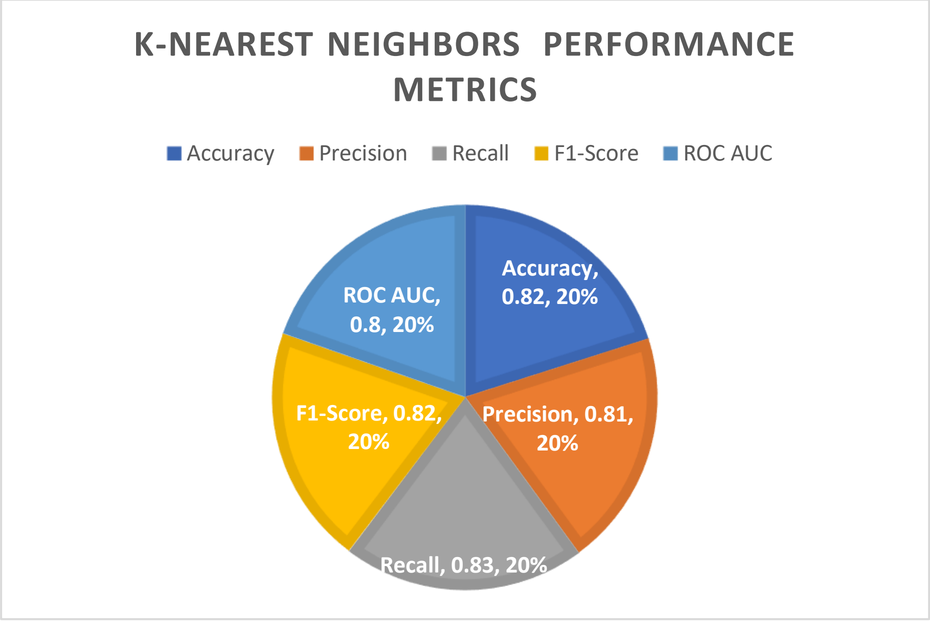

Figure 7: K-Nearest Neighbors performance metrics

Analysis: From figure 7, it is found that in every measure with moderate level of performance KNN is identified. Through recall analysis of 0.83 it means it considers most of the true positives, but lower ROC AUC 0.80 mean less discriminative capability than Random Forest & SVM etc. Proper selection of 'k' and scaling of features may give better result.

vii. Naive Bayes performance metrics

Performance Metrics	Values
Accuracy	0.81
Precision	0.79
Recall	0.83
F1-Score	0.81
ROC AUC	0.77

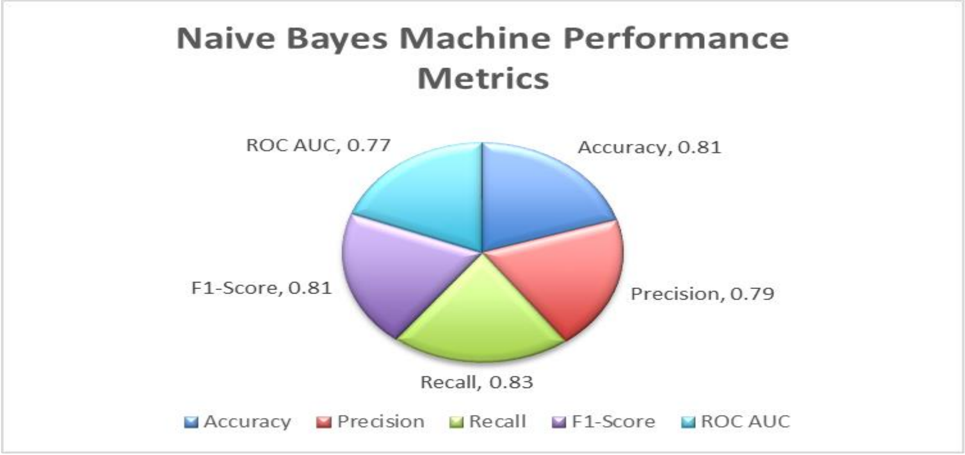

Figure 8: Naive Bayes performance metrics

Analysis: Naive Bayes performs around average, as seen in figure 8, with a recall rate of 0.83 and an accuracy rate of 0.79, indicating that the model is slightly better at TP detection. It displays a ROC AUC of 0.77, which shows less class separation than the other models. This model is good with simple data or simple assumptions on data, but it may not perform very well with the more complex data.

CONCLUSION & RECOMMENDATION

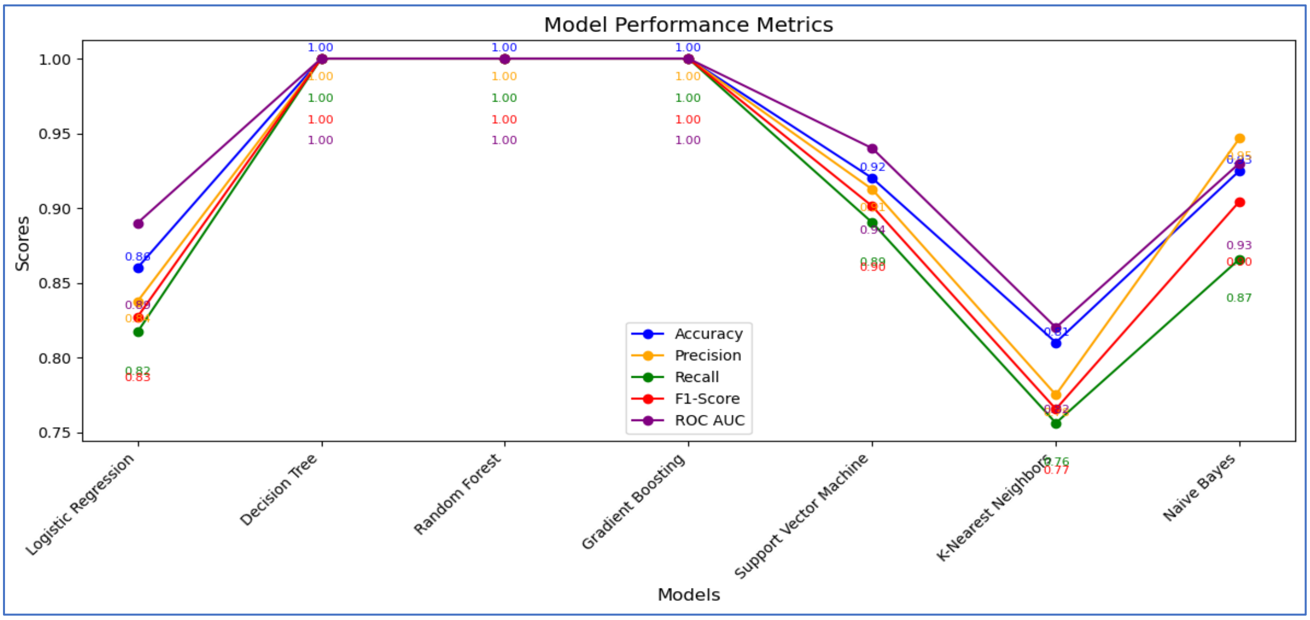

Figure 9: Model Performance Metrics

Figure 9 shows a comparison of the ROC-AUC, Accuracy, Precision, Recall, and F1-Score performance of several ML models. Below are the key insights derived from the results:

Gradual Boosting again shows good performance in each of the parameters, ratings that are at or above 0.90. It strikes a balance between interpretability and performance, often outperforming simpler models like logistic regression, decision tree and random forest. All the metrics scored to perfection (score =1.0) for these models. But this means high predictiveness in the price; meanwhile, it could be over-fitting for small or less-diverged data sets.

Support Vector Machine (SVM):Its Accuracy ranges between 0.89 and 0.92, while its recall, precision and F1-Score vary between 0.89 and 0.92.In terms of all performance measurements, is slightly worse than tree-based and boosting methods.K-Nearest Neighbors (KNN): It has the lowest performance among all models specially in terms

of Precision and Recall below 0.80. This might imply that when dealing with high ,an acceptable checkpoint model characterized by solidity of results obtained across the various measures.

Recommendations:

1. Best Model for Deployment

By the evaluation metrics, Gradient Boosting is suggested to be used in the real world. This model attains high performance, which reduces the value of L(θ), with θ as the model parameters. In order to improve predictions over time, Gradient Boosting minimizes the loss function, which can be the binary cross-entropy in classification problems:

$$L = -1/n * \Sigma[i=1 \text{ to } n] [yi * \log(\hat{y}i) + (1 - yi) * \log(1 - \hat{y}i)]. \tag{18}$$

It owes its proficiency in the prevention of overfitting to basics of overfitting control like the learning rate and pruning of trees.

2. Ensemble Methods

This confirms that the idea of using ensemble methods such as random forest and gradient boosting yields excellent results for predicting heart diseases. These models combine multiple weak learners ht(x) to form a robust prediction:

$$F(x) = \Sigma[t=1 \text{ to } T]\ \alpha t * ht(x), \tag{19}$$

where αt are the learning rate for each learner. These methods are useful in handling massive data in that they can eliminate high variance or biases.

3. Subsequent Assessment and Enhancement

To optimize the model's capabilities:

Optimization of hyperparameters: Modify parameters including learning rate (η) and the number of estimators (T) to enhance performance:

$$F(t+1)(x) = Ft(x) + \eta * ht(x). \quad (20)$$

External testing: Assess the model on independent datasets to confirm generalizability by computing metrics such as the AUC-ROC(area under the receiver operating characteristic curve.)

$$AUC\text{-}ROC = \int[0,1]\, TPR(x)\, dx. \quad (21)$$

TPR denotes the true positive rate.

Feature engineering: Assess the most important variables by applying feature importance techniques like SHAP (Shapley Additive Explanations) values.

$$SHAP(xi) = \varphi 0 + \Sigma[j=1 \text{ to } M]\, \varphi j. \quad (22)$$

These mathematically based methodologies guarantee that Gradient Boosting is resilient, comprehensible, and dependable for practical application in healthcare environments.

REFERENCES

[1] Coffey, Sean, Ross Roberts-Thomson, Alex Brown, Jonathan Carapetis, Mao Chen, Maurice Enriquez-Sarano, Liesl Zühlke, and Bernard D. Prendergast. "Global epidemiology of valvular heart disease." Nature Reviews Cardiology 18, no. 12 (2021): 853-864.

[2] Li, Gloria Hoi-Yee, Ching-Lung Cheung, Albert Kar-Kin Chung, Bernard Man-Yung Cheung, Ian Chi-Kei Wong, Marcella Lei Yee Fok, Philip Chun-Ming Au, and Pak-Chung Sham. "Evaluation of bi-directional causal association between depression and cardiovascular diseases: a Mendelian randomization study." Psychological medicine 52, no. 9 (2022): 1765-1776.

[3] Bhatt, Chintan M., Parth Patel, Tarang Ghetia, and Pier Luigi Mazzeo. "Effective heart disease prediction using machine learning techniques." Algorithms 16, no. 2 (2023): 88.

[4] Naser, Marwah Abdulrazzaq, Aso Ahmed Majeed, Muntadher Alsabah, Taha Raad Al-Shaikhli, and Kawa M. Kaky. "A Review of Machine Learning's Role in Cardiovascular Disease Prediction: Recent Advances and Future Challenges." Algorithms 17, no. 2 (2024): 78.

[5] Kilimci, Zeynep Hilal, Mustafa Yalcin, Ayhan Kucukmanisa, and Amit Kumar Mishra. "Heart Disease Detection using Vision-Based Transformer Models from ECG Images." arXiv preprint arXiv:2310.12630 (2023).

[6] Liu, Aihua, Gerhard-Paul Diller, Philip Moons, Curt J. Daniels, Kathy J. Jenkins, and Ariane Marelli. "Changing epidemiology of congenital heart disease: effect on outcomes and quality of care in adults." Nature Reviews Cardiology 20, no. 2 (2023): 126-137.

[7] Hossain, Md Imran, Ghada Zamzmi, Peter R. Mouton, Md Sirajus Salekin, Yu Sun, and Dmitry Goldgof. "Explainable AI for Medical Data: Current Methods, Limitations, and Future Directions." ACM Computing Surveys (2023).

[8] Dhiyanesh, B., S. Ganapathi Ammal, K. Saranya, and K. E. Narayana. "Advanced Cloud-Based Prediction Models for Cardiovascular Disease: Integrating Machine Learning and Feature Selection Techniques." SN Computer Science 5, no. 5 (2024): 572.

[9] Enad, Huda Ghazi, and Mazin Abed Mohammed. "A review on artificial intelligence and quantum machine learning for heart disease diagnosis: Current techniques, challenges and issues, recent developments, and future directions." Fusion: Pract Appl (FPA) 11, no. 1 (2023): 08-25.

[10] Alam, Md Imran, Haneef Khan, Malik Zaib Alam, Shams Tabrez Siddiqui, Agha Salman Haider, and Mohammad Rafeek Khan. "Preliminary Diagnosis of Diabetes Through Comparative Analysis of Supervised Machine Learning Techniques." In Nanotechnology in Miniaturization, pp. 415-429. Springer, Cham, 2024.

[11] Tsao, Connie W., Aaron W. Aday, Zaid I. Almarzooq, Alvaro Alonso, Andrea Z. Beaton, Marcio S.Bittencourt, Amelia K. Boehme et al. "Heart disease and stroke statistics—2022 update: a report from the American Heart Association." Circulation 145, no. 8 (2022): e153-e639.

[12] Moradi, Hamed, Akram Al-Hourani, Gianmarco Concilia, Farnaz Khoshmanesh, Farhad R. Nezami, Scott Needham, Sara Baratchi, and Khashayar Khoshmanesh. "Recent developments in modeling, imaging, and monitoring of cardiovascular diseases using machine learning." Biophysical Reviews 15, no. 1 (2023): 19-33.

[13] Yılmaz, Rüstem, and Fatma Hilal Yağın. "Early detection of coronary heart disease based on machine learning methods." Medical Records 4, no. 1 (2022): 1-6.

[14] Dubey, Animesh Kumar, Amit Kumar Sinhal, and Richa Sharma. "Impact of machine and deep learning techniques on diseases classification and prediction: a systematic review." International Journal of Advanced Technology and Engineering Exploration 10, no. 106 (2023): 1198.

[15] Qadri, Azam Mehmood, Ali Raza, Kashif Munir, and Mubarak S. Almutairi. "Effective feature engineering technique for heart disease prediction with machine learning." IEEE Access 11 (2023): 56214-56224.

[16] Rababa, Salahaldeen, Asma Yamin, Shuxia Lu, and Ashraf Obaidat. "Predicting Heart Disease and Reducing Survey Time Using Machine Learning Algorithms." arXiv preprint arXiv:2306.00023 (2023).

[17] Mahajan, Palak, Shahadat Uddin, Farshid Hajati, and Mohammad Ali Moni. "Ensemble learning for disease prediction: A review." In Healthcare, vol. 11, no. 12, p. 1808. MDPI, 2023. Rahman, Atta Ur, Yousef Alsenani, Adeel Zafar, Kalim Ullah, Khaled Rabie, and Thokozani Shongwe. "Enhancing heart disease prediction using a self-attention- based transformer model." Scientific Reports 14, no. 1 (2024): 514.

[18] Deepa, Dr R., Vijaya Bhaskar Sadu, and Dr A. Sivasamy. "Early prediction of cardiovascular disease using machine learning: Unveiling risk factors from health records." AIP Advances 14, no. 3 (2024).

[19] Alam, Md Imran, Md Oqail Ahmad, Shams Tabrez Siddiqui, Mohammad Rafeek Khan, Haneef Khan, and Khalid Ali Qidwai. "Blockchain for 5G-Enabled IoHT—A Framework for Secure Healthcare Automation." In Proceedings of Data Analytics and Management: ICDAM 2022, pp.793-801. Singapore: Springer Nature Singapore, 2023.

[20] https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset.