JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Comparative Analysis based on Performance of Classification Algorithms for Heart Disease Prediction

Prity Kumari
PG Student,
Department of computer science,
Shri Ram College of Engineering &Management
Banmore, India

Jitendra Tyagi
Computer Science
Shriram College Of Engineering & Management,
Banmore Gwalior, India

Abstract: *Historical data may be used to forecast future developments and guide businesses in making strategic decisions that provide them a competitive edge, ultimately leading to greater efficiency and more profits. Data from the healthcare business is analyzed by several analysts, who look for patterns that may help them identify and predict diseases that patients and doctors may use in a variety of ways. Examining and assessing heart disease is the primary focus of this study. Worldwide, millions of lives are lost each year due to heart disease (HD). Numerous When heart illness first begins, there are no real-time diagnostic or prognostic techniques available. Using a systematic procedure, I gathered data, cleaned it, selected features using the FCMIM and PCA algorithms, and classified it using machine learning. To evaluate various training and testing data for HD prediction, ML methods are used. Research is being done on a variety of classification methods such as Xgboost, random forest, Extra tree, LGBM, and stacking classifiers. The UCI machinery HD data set is utilized in these experimental investigations. The feasibility of a prediction system was tested using a dataset containing 13 features. Accuracy, Precision, Recall, and the F1-score are used to measure the efficacy of these prediction models. Thus, an additional tree with the highest accuracy is considered the best technique for heart disease prediction based on their experimental research.*

Keywords—Twitter, Sentiment analysis, Machine Learning, LSTM model

I. INTRODUCTION

Researchers are concerned about heart disease (HD); one of the most difficult elements of heart disease is its effective detection & diagnosis within the human body. Even physician experts are unable to detect heart illness with acceptable precision by utilizing early treatments. On the market, there are many

healthcare methods for detecting heart disease. They are both costly and ineffective in predicting the risk of HD in people. Current World Health Organization research suggests that only 67% of cardiovascular problems can be predicted by professionals; hence, there is great potential for studies on human HD. Because of technical improvements, computer science may be employed in medical science.

Heart disease is the main cause of mortality, accounting for approximately one million fatalities per year. Heart disease

accounts for one-third of all mortality, with almost half occurring suddenly and without prior notice. The sole symptom of heart illness is untimely demise. A cardiac attack occurs, If important organs/muscles, such as the heart, fail despite the patient's survival quickly and temporarily. This will lead to serious abnormalities in the heart's components, which may have serious consequences for health, including an increased chance of sudden cardiac arrest. (Singh and Kumar, 2020)

Consistent with present predictions [1], India will be the country with the greatest incidence of heart disease. Heart abnormalities account for single out of 5 deaths in India. One in 3 fatalities will occur during the upcoming 3 years.

There is a variety of HD, all of that causes injury to a distinct group of the heart's inner structures. Thus, any heart condition may be classified as heart disease, and several heart-

related issues will be explored in this study. Coronary HD, also known as coronary artery disease (CAD), is the most prevalent form of HD globally. Having fat buildup in the veins and arteries of the circulatory system is the root cause of this disease. In addition, it stops blood from flowing into the capillaries and veins of the heart, resulting in insufficient oxygen and blood delivery to the organs in the heart (Motarwar et al., 2020),

II. LITERATURE REVIEW

Numerous research has been conducted on predicting HD up to this point. Several data mining and ML algorithms have been implemented and proposed for cardiac patients' records. Depending on the methodology employed, these algorithms have produced a variety of outputs. However, even in this day and age, people are still fighting a losing battle against the numerous complications that come along with having HD. Some recent research studies are as follows:

In [2] for the prediction of HD, an ML-based SCA WKNN (Sine Cosine Weighted K-Nearest Neighbor) algorithm that studies data being kept in the blockchain is presented. Due to the immutability of the data preserved on the blockchain, This technique offers a secure storage space for medical data. Still, it is also a reliable source of educational data. To determine how effective the suggested SCA WKNN is in contrast to other approaches, its efficacy is measured regarding a precision, recall, F-score, root mean square error and accuracy. According to the findings of their analysis, SCA WKNN performs better than K-NN and W K-NN in terms of achieving the highest possible accuracy by a margin of 4.59 percentage points and 15.61 percentage points, respectively. In addition, throughput and latency are tested with peer-to-peer storage and stores based on blockchains. Comparing peer-to-peer storage vs. blockchain-based decentralized storage reveals that the latter has a maximum throughput of 25.03 percent higher.

Rajendran and Karthi, (2022) suggested an innovative machine-learning methodology for HD prediction. High-quality features are provided for improved model efficiency by pre-processing and an entropy-based (feature engineering) FE approach. Cleveland and V data were combined to form the HD dataset. There are fourteen differentiating features that a medical center, Hungarian databases, and Swiss database systems share. In the created dataset on HD, missing values are filled in using the imputation method (IVM), and outliers are removed using the elimination method (OR). These procedures are determined by the degree of correlation between the healthcare attributes and the Mahalanobis distance. Experiment results demonstrated that the IVM + OR pre-processing technique is superior to the other pre-processing procedures utilized for system evaluation. The studies were conducted with the help of several different ML models, with HDD being processed with IVM + OR, ICA (Independent Component Analysis), PCA (Principal Component Analysis), and LDA (Linear Discriminant Analysis) entropy-based approaches. It has been demonstrated that using the suggested entropy-based FE in conjunction with OR and IVM preprocessing results in a significant enhancement across all metrics for the LR and NB classifiers. In addition, the results of the experiments demonstrated that the ensemble model— composed of LR and NB— functioned admirably when run through the suggested pipeline. It had an AUC of 96.8percent, Accuracy of 92.70percent, Specificity of 91.5percent, Precision of 92.50percent, and F1 Score of 0.931, all of which were higher than the results of the state-of-the-art [3].

Sahid *et al.*, (2022) looked into the impact of various techniques for processing imbalanced data on the precision with which HD can be predicted. The majority of the algorithms display superior accuracy while working with balanced data as opposed to imbalanced data. On data that has been balanced utilizing SMOTE Tomek hybrid balancing approaches, SVM, Multilayer Perceptron, and an ensemble of LR & Multilayer Perceptron demonstrate an accuracy of 96 percent. The efficiency of every method was analyzed based on several metrics, including precision, accuracy, f1-score, recall, specificity, cohen kappa, ROC curve & AUC score [4].

In [5], GB (Gradient Boosting Classifiers), RFs, DTs, LRs, and SVMs are just some of the supervised ML approaches utilized on the "UCI ML repository for Statlog (Heart) Data Set" to make predictions about HD. The findings of these algorithms are shown as a bonus, and It is suggested that the method with the best degree of accuracy be used to forecast HD on a web application.

Kumari and Mehta, (2021) used 7 different ML methods for predicting HD and ensemble approaches, such as AdaBoost and voting ensemble method, to enhance the accuracy of algorithms that have performed poorly. Compared to other algorithms, Linear Discriminate Analysis performs admirably: its mean absolute error is 0.185, its mean value is around 0.847, and its false recognition rate is the lowest of any algorithm, at 0.076; however, its accuracy is only coming to around 80percent [6].

Sutedja, (2021) utilized 3 ML models and 3 DL models to obtain the maximum accuracy in forecasting HD. Several ML models, including SVM, LR, and Naive Bayes, are utilized in this study, as well as several DL models, including LSTM (Long Short-Term Memory), CNN (Convolutional Neural Network), and RNN (Recurrent Neural Network). This study found an average accuracy of 86 percent for LR, 88 percent for SVM and 86 percent for Naive Bayes. In the meantime, LSTM reaches 84 percent accuracy, RNN hits 90 percent, and CNN lands at 84 percent. Based on the findings of this study, it can be said that the RNN model is the most effective at forecasting whether a person has HD or not, with an accuracy of 90 percent [7].

Various ML methods are applied in the suggested model, including RF, kNN, and NB, for categorizing HD in a given dataset. Finally, the model is constructed using features and established classification approaches for HD prediction. Between all strategies, the accuracy of the combined feature model and RF methodology is 81 percent [8].

Khurana, Sharma, and Goyal, (2021) assessed multiple supervised learning techniques & feature selection methodologies for HD prediction. On benchmark dataset obtained from Cleveland, UCI ML Repository, the performance of 6 ML classifiers—DT, LR, Naive Bayes, SVM, RF, and k-Nearest Neighbor—as well as 5 feature selection methods, Information Gain, Gain Ratio, One-R, and RELIEF—have been examined. The study's outcomes demonstrate that ML classifiers can predict HD with an accuracy of up to 82.81 percent. The classifier is further enhanced by the feature selection strategies, which increase prediction accuracy to 83.41 percent [9].

In [10], several ML algorithms were trained on the Cleveland HD datasets, including KNN (K-Nearest Neighbors), DT (Decision Tree), LR (Logistic Regression), NB (Naive Bayes), DA (Discriminant Analysis), SVM (Support Vector Machine) and Ensemble. The effectiveness of algorithms was measured utilizing 10-fold cross-validation with and without

Principal Component Analysis. PCA, retaining 9 components, and Ensemble classifiers gave LR the highest accuracy of 85.8 percent. Utilizing a Bagged tree with PCA and maintaining ten elements, the accuracy was 83.8 percent.

III. METHODOLOGY

This section discusses the statement of the problem, research technique, and sequential processes utilized in the study's approach. In addition, a detailed flowchart of the complete research process and an algorithm that has been developed with step-by-step instructions are provided in this section.

A. Problem Statement

The most difficult aspect of HD is its identification. Although there are methods for predicting cardiovascular disease, they are either too expensive or too inaccurate to be of practical use when applied to people. Early identification of heart disorders may reduce deaths and consequences overall. Unfortunately, it is not feasible to precisely monitor patients every day, and doctors cannot consult with patients 24 hours a day since it demands more patience, time, and experience. Now, Because they can get their hands on so much information, they may utilize different machine-learning techniques to uncover hidden trends in the information. With medical data, hidden patterns might be exploited for health diagnosis.

B. Proposed Methodology

Millions of people die every year from cardiovascular disease, according to research conducted all around the world. There is a lack of analysis of the vast volumes of data that healthcare providers collect on heart disease (HD) that might guide important decisions. Diagnosing HD early is crucial now. In terms of global health, heart disease is by far the biggest issue. Detecting this disease early on is vital since it has the potential to protect many lives. Many methods have been tried on the UCI Machine Learning HD dataset. Several researchers have attempted to apply sophisticated methods to this dataset, but there have been no complete analyses of it yet. An ML-based method for predicting cardiovascular disease is the goal of this study. Classification methods, which are integral to prediction, are used to build this model. Model creation makes use of a wide range of classification techniques including Random Forest (RF) and XGBoost (XGB) classifiers. Classifiers are trained and evaluated using data stored in Cleveland's data repository. Additionally, data preparation methods are utilized to verify the size of the data, and information about data, and check null values, and missing values with the aid of an outlier method; this attribute selection strategy, named stander scaler, is subsequently utilized to select crucial characteristics of the input set of data, which either decreases execution time or enhances the performance of the classifier. The machine learning model is the best and most practical alternative for recognizing cardiac disorders, and it may be used in healthcare, where it will play an important part in the cardiology area. The universal language of application was Python.

This section details the techniques and methods that will be used to accomplish the set goals. In addition, this part makes predictions on the outcomes of various methods. What follows are explanations of a few of these techniques:

a) Data Collection

the researcher has utilized datasets on heart disease collected from the UCI ML repository for my analysis here. Just 14 of the 76 traits included in this dataset have been utilized in any of the existing literature. To be more precise, the Cleveland database is the sole one used by ML researchers. Heart disease status is the target for the "goal" field. Its possible values are between 0 (nothing present) and 4 (presence). Most of their experiments with the Cleveland database have focused on determining whether or not it is possible to differentiate between the presence (values 1, 2, 3, 4) and absence (values 0, 2).

Figure 2 demonstrates a count plot for the heart disease input dataset, which can be found in the section below. The existence of a label in the dataset is represented by 0; the absence of a label in the targeted column is represented by 1.

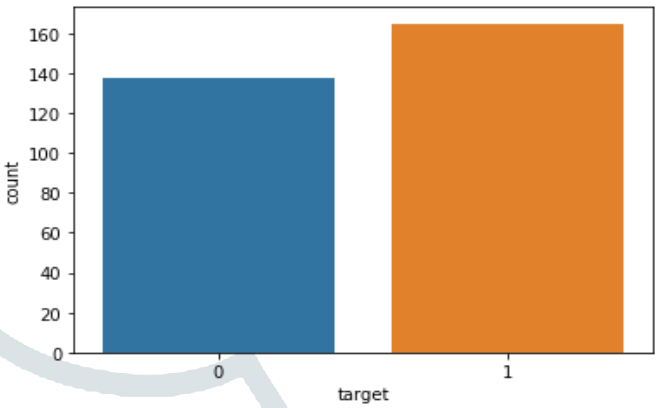

Figure 1: Count plot of the input dataset

b) Data Preprocessing

It is a method of data mining that includes transforming raw data into a predetermined structure. There is a high probability of many inaccuracies in real-world data because of their inconsistency, lack of completeness, or deficiencies in certain behaviors or trends. Preprocessing data is an effective method for resolving a variety of issues. By cleaning and organizing raw data, preprocessing makes it ready for further analysis.

They used data preparation methods such as checking the form of the information, the details of the data (such as datatype, total values, total column, etc.), checking distinct values within the data set (the target column), and using describe after They had collected the dataset (). Inverse () operation (use a data frame to describe information and then transform its format) It uses an outlier approach to check for missing data.

1) Null values

After cleaning up your data, make sure there are no blanks left in your dataset. These missing data reduce the reliability and usefulness of any ML technique. Null values must be removed from a dataset before any ML method can be applied to it.

2) Missing value

When there is no information for some of the expected variables or people involved, this is known as missing values or missing data. There are several potential causes of data loss, including errors in data input, hardware failures, misplaced files, and others. Missing information is commonplace in data sets.

3) StandardScaler

Most ML models need a common scaler to be run as a processing procedure to ensure that the input dataset is uniformly scaled across all relevant functionalities.

c) Feature Extraction

The technique entails picking a feature after the preliminary processing of the data. Important steps in developing a classification model include selecting and extracting features. Effectively, it works by reducing the number of characteristics used to train a classifier, yielding concise models that are computationally demanding yet capable of producing accurate predictions. Moreover, feature selection and extraction techniques are utilized to improve the model's precision.

d) Data splitting

When data is split, it is separated into many groups. When data is split into 2 parts, usually one half is employed to evaluate the model and the other half is used to train it. In this investigation, they split the dataset in half. The data utilized was split into a training set of 70% and a testing set of 30%.

C. Machine Learning Classifier

Machine learning (ML) is a kind of artificial intelligence (AI) that can learn and improve from its own experiences, in addition, to making judgments and predictions. Figure 3 shows the result. The Cleveland dataset is used in this model's training and evaluation processes. First, they use the attributes of the input data to prepare the classifier model; next, The model acquires knowledge from the input data & could identify connections inside the dataset; and, lastly, by testing with new data, the method can forecast the category in which the information relates. By doing so, the system can acquire the ability to learn from data, classify people into healthy and unhealthy categories, and make predictions about cardiovascular disease. I employed the XGB classifier, ET classifier, RF classifier, LGBM classifier, and stacking classifier in this study. The following are some examples of classifiers:

- **XGBoost Model**

It's a powerful and scalable variation on the gradient tree boosting (GB) method. The goal of the GB technique is to decrease the classification model's loss function by the progressive incorporation of weaker learners. This method continuously improves upon previous models by fixing their flaws in the gradient's favor.

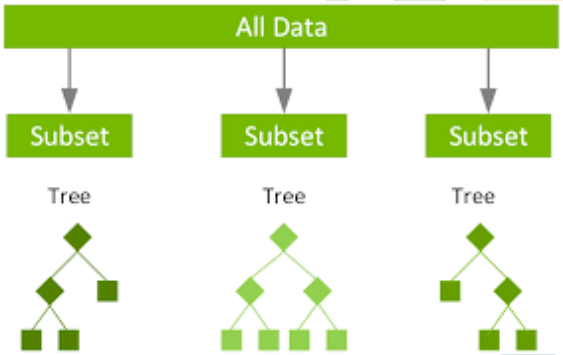

Figure 2: Example of XGBoost Classifier with a working structure

The following are some features of the XGBoost method [11]: (a) the system that is standardized to avoid overfitting problems. (b) A sparsity-aware division method was developed to accommodate the data's diverse sparsity trends. (c) method for dispersed scaled quantile sketching to handle weighted data effectively. (d) Column's in-built construction components provide for simultaneous education (e) Gradient statistics retrieval and storage may both be accomplished using this approach for cache-aware scheduling. (f) blocks used for calculations conducted outside the core.

- **Random Forest**

By creating many decision trees, RF[12] improves the robustness and efficiency of the resulting model. This approach produces a collection of DTs with manageable variation by combining Breiman's bagging sample methods with the random selection of characteristics method [13]. Bagging is used to construct decision trees for each group by randomly replacing some of the training data instances. Each DT within a given group may be used as a "base estimator" to determine an unlabeled example's category. Voting by a large margin is the means through which this is accomplished. Each of the rudimentary DTs models has a vote in determining the predicted category label. The majority-voted-for class label is used to set the instance's category. RF can withstand both noise and overfitting [14].

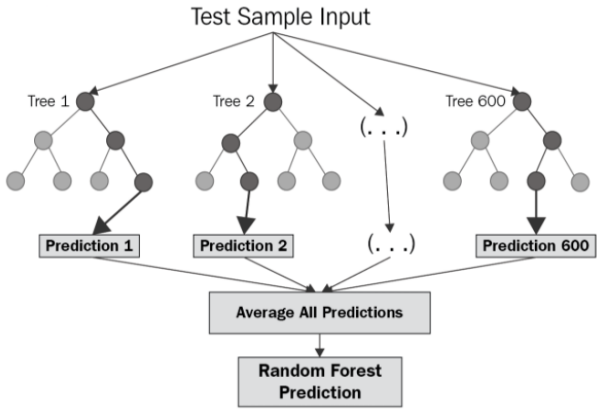

Figure 3: Example of working structure of random forest technique

D. *Proposed method*

Input: Dataset of Heart Disease
Output: Predicated Findings

Step 1: Add the UCI repository's HD Dataset.
Step 2: evaluate the HD dataset that was uploaded

- Null values
- Fill Missing values
- Outlier
- Standard Scaling

Step 3: Select and extract features.
Step 4: Dividing the datasets into 2 stages

- Training (70%) and
- Testing (30%)

Step 5: Deployed ML technique: XGB technique, RF technique, LGBM technique, Stacking technique, ET technique.
Step 6: Compute the F1-Score, Specificity (SPEC), Sensitivity (SENS), Recall (R), Accuracy (ACC), Precision (P), Recall (R), and Error Rate performance metrics.
Step 7: Finally, get predicated results

IV. RESULTS ILLUSTRATIONS

This section provides an overview of the dataset, performance measures, and experimental outcomes. Python programming experiments have been conducted using Jupyter notebook in this suggested work.

A. Dataset Description
B. Performance Metrics

They used several conventional evaluation metrics for medical picture classification, such as accuracy, sensitivity, specificity, recall, precision, & F1-score. They also utilized a CM (confusion matrix) & a graph depicting the relationship between accuracy and the number of training iterations. It is crucial to compare the expected value with the actual value during testing to see whether the model is accurate. This particular research made use of a confusion matrix for its investigation.

The following procedure has been used to evaluate the proposed plan via the use of methods and classification techniques:

1) Accuracy

If the categories of the target variable are almost equal in size, then the accuracy will be rather good. To determine a model's accuracy (ACC), take the sum of all correct predictions (TP + TN) and divide it by the sum of all available data sets (P + N).

$$Acc = \frac{TP + TN}{TP + TN + FN + FP} \ldots\ldots (4.1)$$

2) Precision

As precision evaluates how well results match the target, and recall indicates how many components of the sample were retrieved after being segregated from the rest, it stands to reason that high values for both metrics indicate excellent performance.

$$\text{Precision} = \frac{TN}{TP + FP} \ldots\ldots (4.2)$$

3) Recall

To determine sensitivity, it is possible to divide the number of positive forecasts by the sum of all right predictions. A sensitivity of 1 is the highest possible while a sensitivity of 0 is the lowest. To calculate sensitivities, they utilize the following formula.:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \ldots\ldots (4.3)$$

4) Specificity

The ratio of correct negative predictions to all incorrect ones may be used as a proxy for specificity. A specificity of 1.0 is optimal, whereas a specificity of 0.0 is the worst possible result. they calculate sensitivity using the given formula:

$$\text{Specificity} = \frac{TN}{TN + FP} \ldots\ldots (4.4)$$

5) F1 Score

The F1 Score is obtained by achieving a balance between precision & recall. The F1 score represents a balance between precision and recall. The computation for finding the harmonic mean of a pair of numbers is carried out. Hence, it considers not only the incorrect positive observations but also the incorrect terrible side. The F1 score of a user may be calculated with the help of the following formula:

$$F1 = 2 * \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \ldots\ldots (4.5)$$

C. Experimental Results

In the next section, they show and propose an analysis of the findings obtained making use of the different classification algorithms used to forecast the occurrence of HD. The UCI data collection is being utilized to complete the work of conducting comparative research on the methods. The assessment and selection of the best classification approach have been completed in the course of this study, and the outcomes for the suggested Random Forest (RF) and XGBoost (XGB) techniques are shown below.

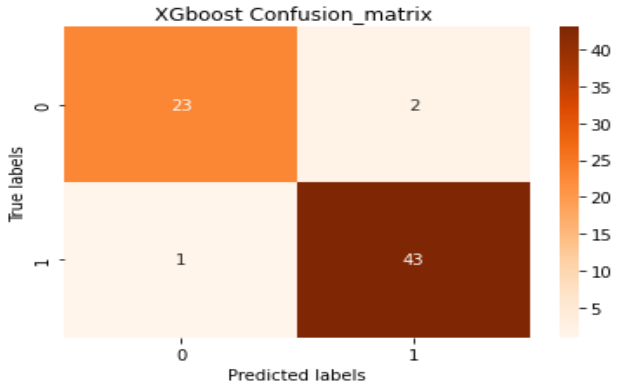

Figure 5: XGBoost technique's Confusion Matrix (CM)

Figure 5 CM of suggested technique. XGB technique has a CM with 23 TP, 43 TN, 1 FP, and 2 FN.

Table 1: Performance measures of XGBoost classifier

Model	Accuracy	F1 score	Precision	Recall	Specificity	Sensitivity
XGBoost	95.6522	96.6292	95.5556	97.7273	97.73	0.92

Table 1 shows the performance measures of the suggested XGBoost classifier. The proposed XGBoost classifier has an accuracy of 95.6522, an F1 score of 96.6292, a precision of 95.5556, a recall of 97.7273, and an error rate of 4.3478.

Table 2: Sensitivity, specificity, and Log loss values for XGBoost classifier

Model	Error Rate	Log-loss metric
XGBoost	4.3478	1.5671

Table 2 shows the sensitivity, specificity, and log-loss values for the proposed XGBoost classifier. The model has specificity, sensitivity, and log-loss values of 0.9733, 0.92, and 1.5671, respectively.

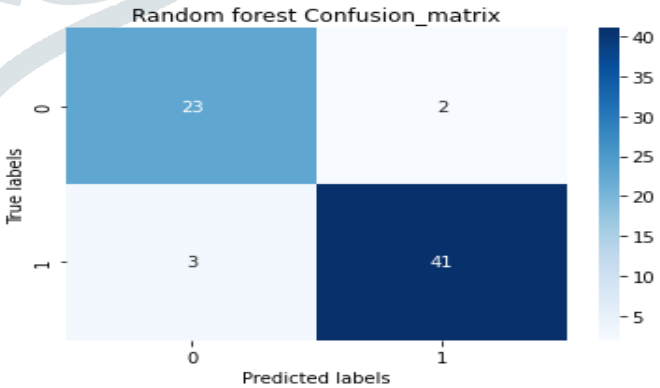

Figure 6: Random Forest technique's Confusion matrix

Figure 6 shows the CM of the presented RF technique. RF Classifier has a CM with 23 TP, 41 TN, 3 FP, and 2 FN.

Table 3: Performance measures of Random Forest

Table 3 shows the performance measures of the proposed RF model. The model has an accuracy of 92.7536, an F1 score of 94.2529, a precision of 95.3488, a recall of 93.1818, and an error rate of 7.2464.

Table 4: Specificity, sensitivity, and log-loss values for Random Forest classifier

Model	Error Rate	Log-loss metric
Random Forest	7.2464	2.6119

Table 4 shows the specificity, sensitivity, and log-loss values of the Random Forest model. The model has specificity, sensitivity, and log-loss values of 0.9318, 0.92, and 2.6119, respectively.

Classification report

	precision	recall	f1-score	support
0	0.88	0.92	0.90	25
1	0.95	0.93	0.94	44
accuracy			0.93	69
macro avg	0.92	0.93	0.92	69
weighted avg	0.93	0.93	0.93	69

Figure 7: Classification report

Figure 7 shows the classification report of the Random Forest classifier. Each row in the table represents a class in the classification problem, and each column represents a metric. The four metrics are accuracy, recall, f1 score, and recall

ROC curve

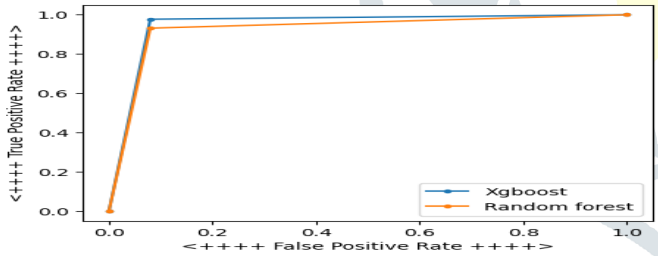

Figure 8: ROC curve

Figure 8 shows the ROC curve for XGB and the random forest model with two axes, Rate of false positives is shown along x, and the rate of correct positives is along y. There are also two lines in it, the blue line is representing XGBoost and the orange line is representing a random forest.

Table 1: Comparison of both proposed models

Model	Accuracy	F1 score	Precision	Recall	Specificity	Sensitivity
XGBoost	95.65	96.62	95.55	97.72	97.73	92
Random forest	92.75	94.25	95.34	93.18	93.18	92

Model	Accuracy	F1 score	Precision	Recall	Specificity	Sensitivity
Random forest	92.7536	94.2529	95.3488	93.1818	93.18	92

Table 1 demonstrates the comparison between the XGBoost and Random Forest classifiers. It evaluates the precision, recall, F1 score, and error rate of the two recommended values.

Table 2 Comparison of Error rate and log loss values

Model	Error Rate	Log-loss
XGBoost	4.3478	1.5671
Random Forest	7.2464	2.6119

Table 2 demonstrates the error rate and log-loss numbers for the XGBoost and Random Forest models presented. XGBoost's error rate and log-loss values are 4.3478 and 1.5671, whereas Random Forest's are 7.2464 and 2.6119, respectively.

V. CONCLUSION

For a long time, The United States' main reason for mortality was cardiovascular disease. Annual global mortality due to cardiovascular disease accounts for around 31%. Until late in the disease process, a patient is sometimes deceived about their symptoms, and some have had difficulty mitigating the risk of heart problems. To prevent cardiac disorders from developing in a large number of patients and to lessen their impact on those already affected, ML techniques have shown to be quite effective. The Cleveland HD Dataset, which can be downloaded from the UCI repository, has 76 characteristics and 303 instances; however, only 14 attributes are used because of missing values. The information was filtered to contain just the most relevant details for ML models using a feature selection technique. Next, the algorithms took over and did the heavy lifting. The XGBoost classifier was the first to be put to the test, and it fared well: it achieved an accuracy of 95.6522, an F1 score of 96.6292, a precision of 95.5556, a recall of 97.7273, a specificity of 97.73, and a sensitivity of 92. The RF classifier was employed as the second approach, and it performed with an accuracy of 92.7536, an F1 score of 94.2529, a precision of 95.3488, a recall of 93.1818, a specificity of 93.18, and a sensitivity of 92. To create reliable forecasts, both algorithms need a substantial quantity of information. Predictions of cardiovascular disease will benefit from this as a consequence of the improved algorithmic precision. From what I've gathered, there are likely to be more upcoming techniques that are even more precise than the aforementioned three.

REFERENCES

[1] V. Sharma, S. Yadav, and M. Gupta, "Heart Disease Prediction using Machine Learning Techniques," 2020. doi: 10.1109/ICACCCN51052.2020.9362842.

[2] H. Hasanova, M. Tufail, U.-J. Baek, J.-T. Park, and M.-S. Kim, "A novel blockchain-enabled heart disease prediction mechanism using machine learning," *Comput. Electr. Eng.*, vol. 101, p. 108086, 2022, doi: https://doi.org/10.1016/j.compeleceng.2022.108086.

[3] R. Rajendran and A. Karthi, "Heart disease prediction using entropy based feature engineering and ensembling of machine learning classifiers," *Expert Syst. Appl.*, vol. 207, p. 117882, 2022, doi:

https://doi.org/10.1016/j.eswa.2022.117882.

[4] M. A. Sahid, M. Hasan, N. Akter, and M. M. R. Tareq, "Effect of Imbalance Data Handling Techniques to Improve the Accuracy of Heart Disease Prediction using Machine Learning and Deep Learning," in *2022 IEEE Region 10 Symposium (TENSYMP)*, 2022, pp. 1–6. doi: 10.1109/TENSYMP54529.2022.9864473.

[5] M. J. Gaikwad, P. S. Asole, and L. S. Bitla, "Effective Study of Machine Learning Algorithms for Heart Disease Prediction," in *2022 2nd International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC)*, 2022, pp. 1–6. doi: 10.1109/PARC52418.2022.9726613.

[6] A. Kumari and A. K. Mehta, "A Novel Approach for Prediction of Heart Disease using Machine Learning Algorithms," 2021. doi: 10.1109/ASIANCON51346.2021.9544544.

[7] I. Sutedja, "Descriptive and Predictive Analysis on Heart Disease with Machine Learning and Deep Learning," 2021. doi: 10.1109/ICORIS52787.2021.9649585.

[8] S. Anbukkarasi, S. Varadhaganapathy, P. Indhiraprakash, V. P. Jeevanantham, and G. Kavin Kumar, "Identification of Heart Disease Using Machine Learning Approach," 2021. doi: 10.1109/ICECA52323.2021.9675865.

[9] P. Khurana, S. Sharma, and A. Goyal, "Heart Disease Diagnosis: Performance Evaluation of Supervised Machine Learning and Feature Selection Techniques," 2021. doi: 10.1109/spin52536.2021.9565963.

[10] K. V. V. Reddy, I. Elamvazuthi, A. A. Aziz, S. Paramasivam, and H. N. Chua, "Heart Disease Risk Prediction using Machine Learning with Principal Component Analysis," 2021. doi: 10.1109/ICIAS49414.2021.9642676.

[11] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," 2016. doi: 10.1145/2939672.2939785.

[12] V. Rohilla, D. S. Chakraborty, and D. R. kumar, "Random Forest with Harmony Search Optimization for Location Based Advertising," *Int. J. Innov. Technol. Explor. Eng.*, 2019, doi: 10.35940/ijitee.i7761.078919.

[13] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[14] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *Stata J.*, 2020, doi: 10.1177/1536867X20909688.