[1]Mrs. S. Madhumalar,
Dr. S. Sivakumar

Optimizing Diabetic Coronary Heart Disease Prediction Models Using Ensemble Learning Approaches

JES Journal of Electrical Systems

Abstract; Diabetic Coronary Heart Disease (DCHD) continues to be a global threat, taking millions of lives every year. The potential of healthcare data on heart disease to influence decisions is still largely unrealized, despite the data being available in massive quantities. The identification of cardiovascular conditions, such as heart attacks and coronary artery disorders, presents a formidable obstacle that traditional clinical data analysis finds difficult to overcome. This work aims to improve the accuracy of weak classification algorithms and demonstrate the usefulness of the method in early diabetic heart disease prediction by implementing it on a medical dataset. In response, this research introduces a novel heterogeneous ensemble learning approaches (ELA), to forecast cardiac disease early by utilizing a novel combination comprising four base classifiers adaptive boosting (AdaBoost), K-Nearest Neighbour (KNN), Random Forest (RF), Support Vector Machine (SVM) to improve the forecast results' accuracy. The two-feature selection wrapping approaches in this article are backward elimination and forward selection. Proposed method is compared with artificial neural networks (ANN), support vector machines (SVM), decision tree (DT), Naive bayes (NB) and eXtreme Gradient Boosting (XGBoost). Evaluation metrics, such as F1 score, recall, accuracy, precision, are employed to assess the proposed method. The experimental results surpassed other methods with classification accuracies of 91% and 92% on the Cleveland and Framingham datasets, respectively. Additionally, the ROC curve analysis showed that the ensemble method had higher ROC (Receiver Operating Characteristic) values, indicating better performance compared to other methods. The results demonstrated that the challenge of classifying an unbalanced diabetic heart disease dataset might be solved using an ensemble-learning framework based on other models.

Keywords: K-Nearest Neighbour, Diabetic Coronary heart disease, Random Forest, Ensemble Learning.

1. INTRODUCTION

DCHD is regarded as one of the risks that people worldwide face. According to data from various international healthcare organizations, cardiovascular illnesses claimed 17.9 million lives in 2019 (32% of all deaths worldwide); by 2030, the number is predicted by WHO [20] to rise to 23 million. In order to better forecast heart-related disorders and take preventative steps in advance, computational intelligence techniques must be explored. Thanks to significant developments in big data, technology, and the storage, acquisition, and recovery of information, artificial intelligence (AI) is becoming more and more prevalent in the field of cardiology (Jahnson et al., 2018).

Additionally, a great deal of research may be done on machine learning (ML) techniques to support healthcare governance and resources for improved patient health services. Hospital administration, telemedicine platforms, practitioners, healthcare professionals, and patient categories will all immediately profit from this. A widely utilized set of machine learning algorithms, together with its modifications, is used to forecast heart failure in the cataloging of genetic heart disorders and control individuals. Alom and associates (2021) DCHD prediction algorithms include KNN, DT, SVC, LR, and RF machine algorithms (Gour et al., 2022; Juhola et al., 2022). Three categories can be used to group machine learning techniques (2022): Unsupervised ML: data-driven (clustering); Reinforcement Learning: learning from errors (gaming); Supervised ML: task-driven (classification/regression).

In this study, ML classifiers such as ANN, SVM, DT, NB and XGBoost are used to assess the accuracy of various models and demonstrate how they can predict the presence of Coronary heart disease. This study develops a model for improved diabetic heart disease prediction through the application of ensemble learning approaches (ELA) methodologies. To get the best results for predicting cardiac illness, an ELA with four machine learning (ML) algorithms is used: AdaBoost, SVM, KNN and RF.

[1]Assistant Professor, Cardamom Planters' Association College,Affiliated to Madurai Kamaraj University, Madurai
[2]Principal, Cardamom Planters' Association College, Affiliated to Madurai Kamaraj University ,Madurai

The key contributions of proposed work are listed as follows:

- Ensemble method based on extreme adaptive boosting (AdaBoost) for Diabetic Coronary heart disease prediction is proposed.
- The suggested meta-heuristics-based feature selection strategy and a variety of data pre-processing techniques are used to handle missing data values and lower the dataset's dimensionality and complexity.
- In this work, two popular datasets on heart disease were pooled to evaluate the effectiveness of the model.
- Evaluation metrics, such as F1 score, recall, accuracy, precision and ROC are employed to assess the suggested method.
- Finally, when compared to state-of-the-art accuracy, the suggested ensemble learning algorithms provide superior accuracy and ROC curves shows, ensemble methods exhibit superior ROC and performance, outperforming other methods.

The paper's remaining content is organized as follows: Section 2 reviews recent relevant state-of-art research. Section 3 describes proposed ensemble learning approach Section 4, ensemble learning approaches is tested, compared with alternative methods. Lastly, Section 5 summarizes the findings and scope of the study.

2. RELATED WORKS

This section presents the literature on the use of ML and DL algorithms in the prediction of diabetic cardiac heart disease. In addition to using a clustering technique, Maini E. et al.(2018) offered unsupervised learning for classifying the Cleveland Dataset. They created a range of models for the feature selection method of diagnosing heart sickness, such as DT, LR, RF, naive Bayes, and SVM. They saw that models including MATLAB, Weka for quick evaluation. Rapid Miner's accuracy has increased. The method with the best accuracy, 90.74%, was used. Saboor A. et al.(2022) employed nine ML classifiers, including AB: AdaBoost (Adaptive Boosting), LR: Logistic Regression, ET: Extra Trees (or Extremely Randomized Trees), MNB: Multinomial Naive Bayes,CART: Classification and Regression Trees, SVM: Support Vector Machine, LDA: Linear Discriminant Analysis, RF: Random Forest, XGB: XGBoost (Extreme Gradient Boosting), on dataset, Both prior to and following adjusting the hyperparameters. The diagnosis of heart illness was analyzed using a variety of ML, such as SVM, which had an accuracy of 96.2%. Tama et al. introduced a unique ML-based method for detecting coronary heart disease (CHD), called classifier ensembles byMarimuthu et al. (2018)**.** This led to the creation of a two-tier ensemble, where one ensemble was formed on top of specific ensemble classifiers. Several datasets(Clevand,Fragmingham)related to heart disease were used to evaluate the model, and the suggested method outperformed all base classifiers in the ensemble.

To forecast cardiopathy, Modepalli et al. (2021) devised a novel machine-learning (ML) strategy. Regression and classification are two data processing techniques that were used on the Cleveland cardiopathy dataset. RF, DT and Hybrid Model ML techniques are used in the implementation. According to output of the trial, the hybrid model has an overall accuracy of 88.7% in predicting coronary heart disease (CHD). Yadav and Pa (2020) assessed the accuracy, precision, and sensitivity of four algorithms they devised for using trees to categorize data. The experimental settings utilized for the analysis used M5P, RT, and RF ensemble techniques.

Tuncer et al.(2022) for ensemble approach an ECG signal identification method including preprocessing, feature extraction, concatenation, selection, and classification was presented. ECG signals were divided into fifteen subbands during the preprocessing. Z-Alizadeh Sani dataset (2018) was used to evaluate the feature importance, and synthetic minority oversampling was used to balance the five features that Velusamy and Ramasamy (2021) selected for their ensemble approach. The weighted average voting (WAVEn) method identified CAD with 100% accuracy and specificity when applied to the balanced dataset. Kataria, R. et al. (2021) highlight heart disease and associated dangers and present machine learning algorithms. They predicted cardiovascular disease using these machine learning techniques, and they also provided a comparison of supervised learning models used during diagnosis process. Compared to other classifiers, the Logistic Regression classifier in this suggested model exhibits a superior accuracy of 93.40%.

A supervised learning classification was used by Shah, D. et al. (2020) on an existing dataset of cardiovascular disease patients from the Cleveland resource of the UCI repository. There are 76 options and 303 components in the set. Analysis takes into account precisely 14 of the 76 attributes, which is crucial for illustrating how learning approaches function. KNN reaches accuracy (89.98%) with lower mistake value, as demonstrated by this paper. Golande and pavan (2019) suggested an architecture that entails testing several

algorithms and preparing input data prior to training. The author suggests using Adaboost since it enhances how all ML methods are presented. The author also backed the notion of adjusting settings to obtain great precision. Using the UCI dataset, researchers proposed a deep learning approach for the investigation and diagnosis of cardiac disease by Sharma and parmar (2020). The KNN, RF, SVM, and DT algorithms were examined as potential machine learning models that can accurately, precisely, and with high recall anticipate cardiac disease. Forecasting model use UCI machine learning archives for cardiac illnesses, classification generated by SVM yielded the greatest accuracy of 86% by Arunpradeep and niranjana (2020).

3. MATERIALS AND METHODS

Computational complexity is crucial, particularly when dealing with big datasets. Five steps make up the suggested framework: i) Data Collection and Pre-processing ii) Feature selection iii) Classification iv) Training models and v) Evaluating ensemble model.

3.1 Dataset

The Cleveland dataset [18] from University of California, Irvine repository and Framingham dataset from Kaggle website [6] were used for the tests. Cleveland heart dataset contains 303 instances and 14 attributes in the collection. Six numerical attributes and eight categorical attributes are present. Table 1 displays the Cleveland dataset description. The data originate from the well-known Framingham Heart Study, which was originally designed as a 20-year cohort study of people living in Framingham, Massachusetts, between the ages of 30-59 in 1948. The observations in the current data pertain to n = 1,363 individuals. Table 2 displays the Framingham dataset description. Age, Sex, Cholesterol, Blood pressure, Blood sugar (Glucose), and Heart rate features are selected from two datasets for performance evaluation. This ensemble method suits real-time datasets, enhancing adaptability and accuracy in dynamic environments. The quality of training data is a critical factor in determining the effectiveness of any machine learning system. In order to preserve data consistency before conducting any analysis, data preparation is used as one of the most important steps. Making the data appropriate for using machine learning algorithms is the first step in this process.

Table 1: Cleveland heart dataset

Variable Name	Role	Type	Description	Units	Missing Values
Age	Feature	Continuous			no
Sex	Feature	Binary			no
chest-pain	Feature	Categorical	chest pain type		no
rest-bp	Feature	Continuous	resting blood pressure		no
serum-chol	Feature	Continuous	serum cholesterol	mg/dl	no
fasting-blood-sugar	Feature	Binary	fasting blood sugar > 120 mg/dl		no
electrocardiographic	Feature	Categorical	resting electrocardiographic results		no
max-heart-rate	Feature	Continuous	maximum heart rate achieved		no
angina	Feature	Binary	exercise induced anigna		no
oldpeak	Feature	Continuous	oldpeak = ST depression induced by exercise relative to rest		no
slope	Feature	Integer	the slope of the peak exercise ST sgment		no
major-vessels	Feature	Continuous	number of major vessels (0-3) colored by fluorosopy		no
Thal	Feature	Categorical	thal: 3 = normal; 6 = fixed defect; 7 = reversable defect		no
heart-disease	Target	Integer			no

Table 2: Framingham heart dataset

Variable	Description
Age	Age in years (32 to 70)
Male	Gender instance (1 = Female, 0= Male)
Education	Level of education (1 to 4)
CurrentSmoker	Whether or not the patient is a current smoker 0 : no 1 : yes
CurrentSmoker	Whether or not the patient is a current smoker 0 : no 1 : yes
CigsPerDay	The number of cigarettes that the person smoked on average in one day
BPMeds	Whether or not the patient was on blood pressure medication 0 : no 1 : yes
PrevalentStroke	Whether or not the patient was on blood pressure medication 0 : no 1 : yes
PrevalentHyp	Whether or not the patient was hypertensive 0 : no 1 : yes
Diabetes	Whether or not the patient had diabetes 0 : no 1 : yes
TotChol	Total cholesterol level
SysBP	Systolic blood pressure
DiaBP	Diastolic blood pressure
BMI	Body Mass Index
Heart Rate	Measure of heart rate
Glucose	Glucose level
TenyearHeart	whether or not the patient will develop heart disease in the future ten years (target) 0 : no 1 : yes

3.2 Data Collection and Pre-processing

Cleaning: Usually, the values obtained have noise and erroneous numbers. To produce output that is precise and effective, such data needs to be cleaned up of noise and incompleteness.

Transformation: This is the process of changing how information is presented from one form to another to improve comprehension. There are techniques for aggregation, and smoothing used.

Integration: If authors don't combine material from several sources that readers could be hard for us to understand what they're reading.

Reduction: To get pertinent results, the complicated data that has been gathered needs to be arranged. To enable reliable recognition, the information is first categorized and divided into training and testing sets. For each set, a number of approaches are applied.

3.3 Feature selection

Getting a subset of highly significant features is the main goal of feature selection, which also helps to shorten the machine learning classifiers' training time. Because irrelevant characteristics frequently have an impact on the machine learning classifier's classification efficiency, identifying the best features is an essential first step.

3.4 Data partitioning

Using the chosen characteristics, the Cleveland and Framingham datasets are split at random into training and testing subsets for the creation of the four basic classifiers. In Dataset, 20 percent is used for testing and 80 percent of it is used for training. During the training phase, each base classifier is adjusted based on the training errors using 10-fold cross-validation with three repetitions and an appropriate measure. The averages of the results are then calculated using equation 1. The original dataset D, which may be divided into several homogenous sets, should be regarded as the root node as it denotes whole population. Choose a data point at random from variable set consisting of j = 1, 2,..N. and i = 1, 2,... N. To allow for further choices, mean-based partitioning approach, divide D into two parts:

$$D = \begin{cases} D_{11}, & if\ x_{ij} < \bar{x}_j \\ D_{12}, & if\ x_{ij} \geq \bar{x}_j \end{cases} \quad (1)$$

Viewing D11 and D12 as root nodes and applying (1) means looking at each child node independently.

3.5 Classification Models

Using an existing data, classification is a supervised learning process that forecasts the result. In order to increase classification accuracy, an ensemble of classifiers is suggested as a method in this work for diagnosing diabetic heart disease using classification algorithms. Using the test dataset, the classifiers' efficiency is evaluated. Using the Cleveland and Framingham datasets, adaptive boosting combined with KNN, RF, and SVM can be utilized to predict DCHD. This dataset, which is used to train and evaluate the methods depicted in Figure 1, includes a variety of clinical and administrative patient data.

3.5.1 Support Vector Machine

One of the most well-known and useful methods for addressing problems with data classification, learning, and prediction is support vector machines (SVM). Support vectors are data points that are near to decision. In infinite dimensional space, it classifies data vectors using a hyper plane. The maximal margin classifier, which helps identify fundamental linearly separable training data classification problem using binary classification, is basic type of SVM. In real-world complexity, the hyper plane with the largest margin is identified by the maximal margin classifier. SVMs use a range of kernel techniques. Equation (2) defines a mathematical description of value factor for SVM.

$$(\theta) = \frac{1}{2}\sum_{j=1}^{n} \theta^2 \, j \quad (2)$$

3.5.2 Random Forest

Regression analysis and classification both use random-forest (RF) techniques. A tree-based depiction of the data is used to make forecasts. RF technique may produce identical results when used to big datasets, even in cases where a sizable part of record entries are missing. RF composed of 2 stages: the first is RF creation; the second is projection using the initial step's established classification which is defined in equation (3).

$$Gini = \; 1 - \sum_{i=1}^{n} (p_i)^2 \quad (3)$$

The likelihood of an object falling into a particular class or feature is represented by the value of p_i.

3.5.3 K-Nearest Neighbour

K-NN is an easy method of using similarity metrics to categorize fresh instances. To determine who the closest neighbour in k-NN is, it can compute the distance between two data points in a number of different ways. But because all of the training data must be kept in memory, the k-NN uses more memory. However, k-NN is widely employed in medical diagnostics, particularly for the identification of heart problems. A variety of distance measurement systems are supported by the k-NN. However, k-NN indicates that the k rows that are selected have small distances to the target instance. Distance is typically used to determine similarity which is defined in equation (4)

$$d(x,y) = \sqrt{\sum_{i=1}^{n} ((y_i - x_i))^2} \quad (4)$$

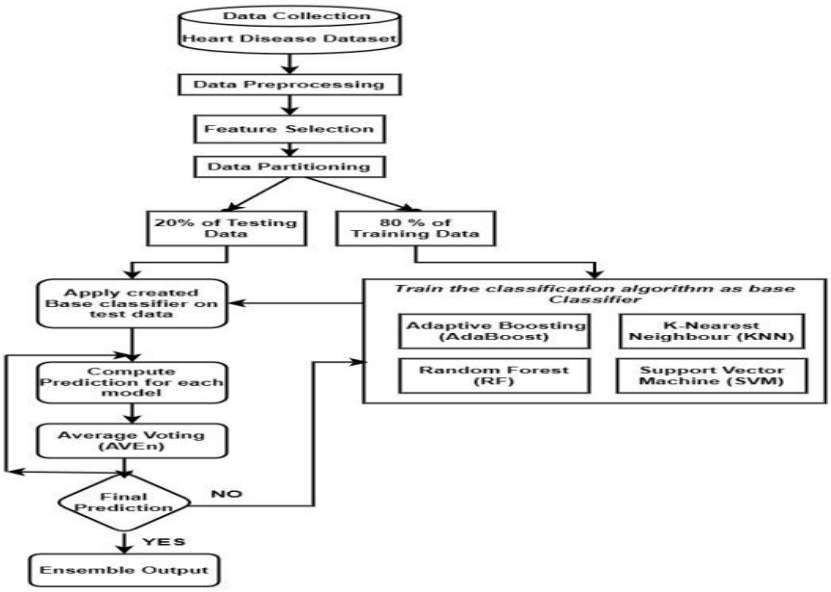

Figure 1: Architecture of proposed AB-DCHD Ensemble Model

3.5.4 Adaptive Boosting

Adaptive boosting, also referred to as adaboost, is a method of data classification that pools the knowledge of multiple weak learners. The cases that were handled correctly will be given less weight, and the ones that were incorrectly classified will be given more weight, by altering the weights assigned to each instance. By integrating multiple weak learners, or simplistic models, it produces a stronger classifier overall.

The approach first computes the error, then utilizes the data to train the first weak learner. The outcome of the samples that were misclassified are given more weight and are instructed with greater focus on the poor learner that comes after. There are several iterations of this process. Prior to being merged to create the final model, each weak learner's prediction is assigned a weight corresponding to its accuracy. A training set of m instances and r attributes is assumed by Adaboost. $TR = \{(tr_1^1, tr_1^2, \dots\dots, tr_1^r), (tr_2^1, tr_2^2, \dots\dots, tr_2^r), \dots, (tr_m^1, tr_m^2, \dots\dots, tr_m^r)\}$. Total weak classifiers is represented as F = $\{F_1, F_2, F_3, \dots\dots, F_{r-1}\}$, $F_{k(tr)} \in \{-1,1\}$, loss function is shown in equation (5)

$$L(F_k(tr), tr^r) = \begin{cases} 0, if\ F_k(tr_i) = tr_i^r \\ 1, if\ F_k(tr_i) \neq tr_i^r \end{cases} \quad (5)$$

Linear combination of the classification output supplied by F is the ultimate result of Adaboost. Equation (6) provides the definition of Adaboost's expected output.

$$L(TR) = sign\left(\sum_{j=1}^{r=1} E_j F_j(tr)\right) \quad (6)$$

It should be mentioned that characteristics in the training dataset aside from the class attribute is what determines the number of weak learners F. Given that there are r attributes in TR, which includes class attributes, there are r − 1 weak learners overall.

Algorithm 1:Adaboost Ensemble

Input

$TR: = \{(tr_1^1, tr_1^2, \dots\dots, tr_1^r), (tr_2^1, tr_2^2, \dots\dots, tr_2^r), \dots, (tr_m^1, tr_m^2, \dots\dots, tr_m^r)\}$, Training scenarios in dimensions of m x r

$TS := \{(ts_1^1, ts_1^2, \dots\dots, ts_1^r), (ts_2^1, ts_2^2, \dots\dots, ts_2^r), \dots, (ts_m^1, ts_m^2, \dots\dots, ts_m^r)\}$, testing instances with n x r dimension

Output

Predicted class labels are filled in attribute t_n^r in the TS.

Process

begin

for i: = 1 to m

$w_i^1 := 1$

end // *Training phase*

for j: = 1 to r – 1

$$\varepsilon_j = \frac{\sum_{i=1}^{m} W_i^j I(F_j(tr_i) \neq tr_i^r}{\sum_i^m W_i^j}$$

$E_j := \log\frac{1-\epsilon_j}{\epsilon_j}$ // *performance of the stump*

for i: = 1 to m

$w_i^{j+1} := w_i^j e^{E_j I(F_j(tr_i) \neq tr_i^r}$ // *Weight updated*

end

end // *Testing phase*

for i: = 1 to n

$$ts_i^r = sign\left(\sum_{j=1}^{r-1} E_j F_j(ts_i)\right)$$

end
return TS
end

3.6 Average Voting Ensemble (AVEn)

Average Voting Ensemble combines predictions from multiple models by averaging their outputs to make a final decision. It accomplishes this by simply combining the chances in the past that the M classification models yielded for each sample. Sample X is HD if $\hat{y}$ > 0.5; else, it is Normal.

$$\hat{y} = \frac{1}{M}\sum \Pr\left(CA_1(X)\right) + \Pr\left(CA_2(X)\right) + \cdots . + \Pr\left(CA_M(X)\right) \quad (7)$$

4. EXPERIMENTAL RESULTS

4.1 Performance Evaluation

Performance Evaluation is crucial for verifying the accuracy of the proposed method. The ensemble method is compared with ANN, SVM, DT, NB, and XGBoost for comprehensive analysis.

4.2 Evaluation Metrics

Four key factors are used to assess the success of the suggested framework: TP (true positive) indicates the proportion of true cases that are also classified in the true category; TN (true negative) indicates the proportion of false cases that are classified as false; FP(false positive) indicates the proportion of false cases that are classified as true; and FN (false negative)indicates the proportion of true cases that are classified as false. The following is a discussion of the various evaluation criteria employed in the current study.

Accuracy

It evaluates how well the model can forecast the future. Out of all the categorization predictions, it counts the number of accurate predictions. Accuracy is computed using the confusion matrix in the following way:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Recall

Recall, sometimes referred to as sensitivity, and evaluates the model's capacity for producing precise, affirmative predictions. It determines how many individuals with diabetic heart disease the classifier correctly predicted. The formula for recall is as follows:

$$\text{Sensitivity / Recall} = \frac{TP}{TP+FN}$$

Precision

Counting the number of patients who have been correctly predicted to have diabetic heart disease out of the actual patients is called precision. It assesses the model's capacity to generate solely pertinent outcomes.

$$\text{Precision} = \frac{TP}{TP + FP}$$

F-score

By calculating the harmonic mean of recall and precision, it is measured. When taking into account other statistical variables, the F1-score is computed as follows:

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

4.3 Comparison with existing work

The proposed Ensemble method is evaluated with Cleveland and Framingham datasets shown in Figure 2 and 3. Tables 3 and 4 compare the suggested classification algorithm's performance with that of other currently in use.

Ta

ble 3: Accuracy, Precision, Recall, F1-Score analysis on Cleveland

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
SVM	75.2	62.1	61.4	64.3
DT	80.6	67.5	63.7	71.6
ANN	79.1	77.9	75.6	69.4
NB	85.4	69.2	70.2	77.8
XGBoost	89.7	78.6	77.8	80.1
Ensemble Method	91.8	80.9	79.3	84.9

The Cleveland dataset's heart disease categorization findings demonstrate that different models perform and exhibit accuracy at different levels. SVM obtained a 75.2% accuracy rate, with 61% precision, recall, and F1 scores. The accuracy of the ANN was 79.1%, and its precision, recall, and F1 scores were all 73%. When evaluated, decision trees showed a marginally higher accuracy of 80.6% along with comparable precision, recall, and F1 scores. When compared to other methods, XGBoost's accuracy was marginally higher at 89.7%, and its Precision, recall, and F1 scores were comparable. However, NB outperformed them both, attaining an accuracy of 85.4% and a well-balanced 71% F1 score, recall, and precision. With the help of four base classifiers—AdaBoot, KNN, RF, and SVM—the Ensemble Learning Approach performed exceptionally well, attaining an astounding 91.8% accuracy as well as equally distributed precision, recall, and an F1 score of 81%. This suggests that diabetic heart disease prediction can be greatly improved by using ensemble models.

Table 4: Accuracy, Precision, Recall, F1-Score analysis on Framingham

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
SVM	77.5	64.3	68.3	78.7
DT	79.2	67.6	73.4	69.5
ANN	86.5	70.4	71.5	75.9
NB	81.9	75.2	78.8	81.7
XGBoost	84.7	78.9	82.6	79.2
Ensemble Method	92.3	85.4	83.7	88.8

The Fragmingham dataset's heart disease categorization findings demonstrate that different models perform and exhibit accuracy at different levels. SVM obtained a 77.5% accuracy rate, with 64.3% precision, recall, and F1 scores. The accuracy of the ANN was 86.5%, and its precision, recall, and F1 scores were all 75.9%. When evaluated, decision trees showed a marginally higher accuracy of 79.2% along with comparable precision, recall, and F1 scores. When compared to other methods, XGBoost's accuracy was marginally higher at 84.7%, and its Precision, recall, and F1 scores were comparable. However, NB outperformed them both, attaining an accuracy of 81.9% and a well-balanced 81.7% F1 score, recall, and precision. With the help of four base classifiers—AdaBoot, KNN, RF, and SVM—the Ensemble Learning Approach performed exceptionally well, attaining an astounding 92.3% accuracy as well as equally distributed precision, recall, and an F1 score of 88.8%. This suggests that diabetic coronary heart disease prediction can be greatly improved by using ensemble models.

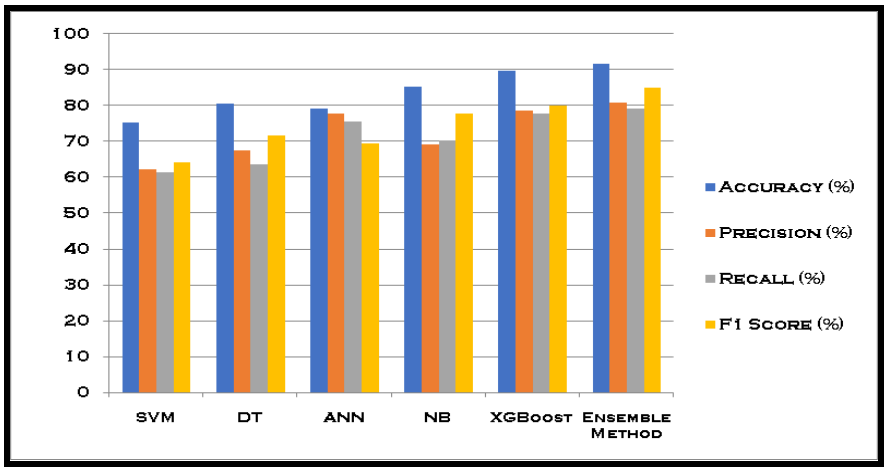

Figure 2: Comparison of proposed method AB-DCHD with other models on Cleveland dataset

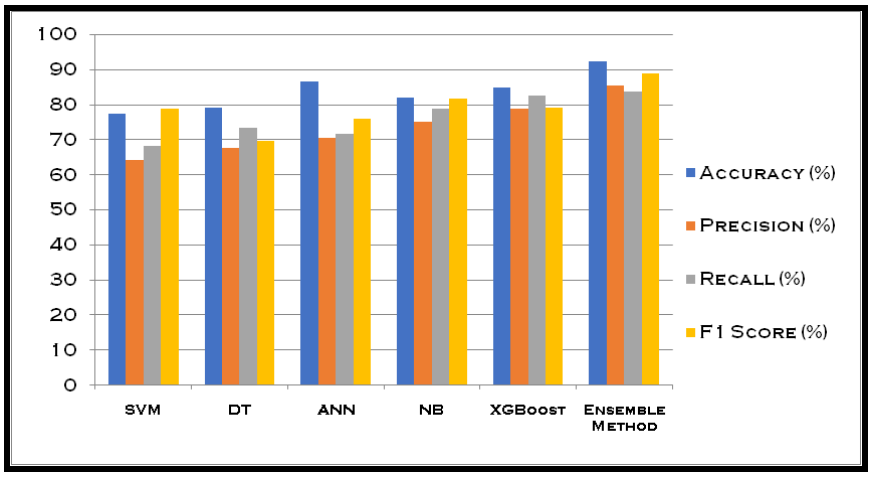

Figure 3: Comparison of proposed method AB-DCHD with other models on Framingham dataset

Another method for evaluating machine learning models is the receiver operating curve (ROC), which is used to calculate percentage or degree to which the proposed model can distinguish between distinct classes of dataset instances. Figure 4, 5 illustrates the ROC curve for proposed ensemble model. The presentation estimation for grouping problems at various edge settings is the AUC-ROC curve. The link between the machine learning model's sensitivity and specificity is displayed via ROC. The degree or proportion of distinctness is addressed by AUC. It indicates the degree to which the model is capable of class recognition. Plotting the True Positive Rate(TPR) against theFalse Positive Rate(FPR), with TPR on the y-pivot and FPR on the x-pivot, yields the ROC curve.

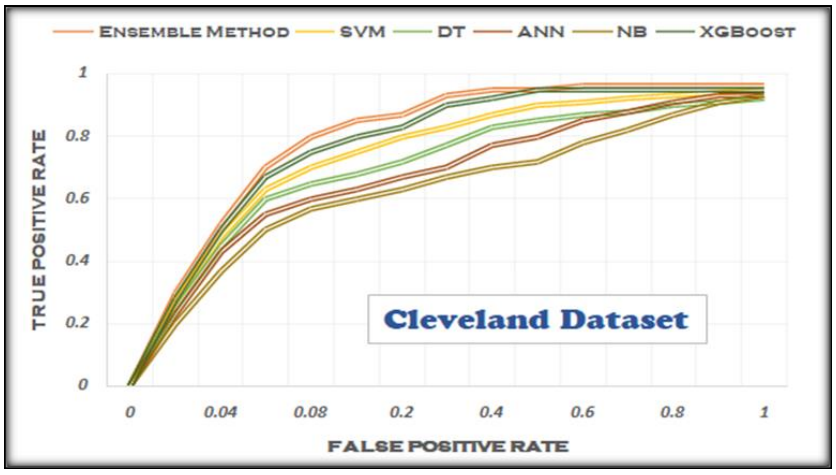

Figure 4.ROC curve for proposed framework with existing models using Cleveland dataset

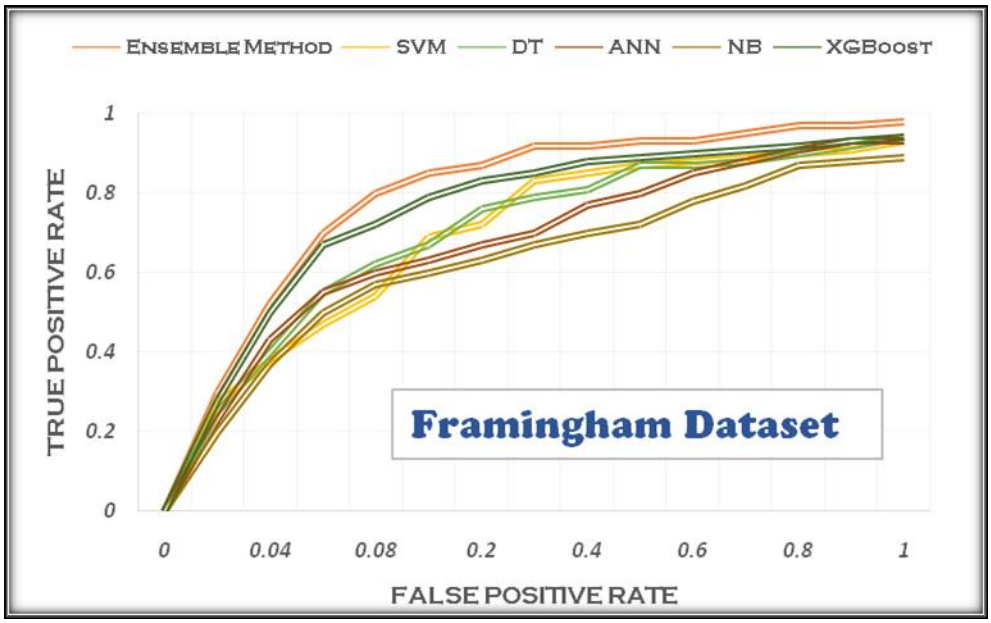

Figure 5.ROC curve for proposed framework with existing models using Framingham dataset

5. CONCLUSION AND FUTURE WORK

Ensemble Learning techniques have recently demonstrated outstanding performance in a wide range of applications and have garnered a lot of attention. These techniques can improve a single classifier's capacity for generalization. The techniques covered have been chosen because they are widely used in a variety of machine learning applications. This work aimed to fill a vacuum in the literature by providing a clear and simple explanation of the chosen ensemble learning algorithm. This study examines machine learning's efficacy in predicting diabetic coronary heart disease outcomes. The most popular Cleveland heart disease dataset and Framingham dataset is used for evaluation. Proposed ensemble method is applicable for real-time datasets, enhancing their utility and performance.

The accuracy of the base models was increased by the bagging, boosting, and stacking ensemble approaches. Ensemble methods boosting and bagging with feature extraction algorithms KNN, RF, SVM and adaptive boosting are used to improve predicting heart disease performance. With an average accuracy of 91%, the proposed model beat every other model in the current study, according to a comparative analysis of the Cleveland dataset. Furthermore, the statistical outcomes of recall (79.3%), F-score (84.9%), and precision (80.9%) demonstrate the efficacy of the current methodology.

Future research will focus on using ensemble learning to address big data-related problems. Big data has attracted a lot of interest lately. For processing and modeling large amounts of data, deep learning has shown to be a useful technique. In contrast to conventional machine learning models, deep learning models tend to be more intricate and challenging to train. It may be beneficial to investigate the applicability and advantages of ensemble methods in this rapidly expanding field. Furthermore, the use of ensemble learning with big data has already been the subject of a few research projects. More study centered on ensemble learning, though, might be beneficial in this field.

REFERENCES

[1] Abdul Saboor, Usman Muhammad, Ali Sikandar, Ali Samad, Faisal Abrar Muhmmad, Ullah Najeeb. A method for improving prediction of human heart disease using machine learning algorithms. Mobile Inf Syst 2022;2022: 1410169. 9 pages.

[2] Alizadehsani, Z., Alizadehsani, R., &Roshanzamir, M.(2018) [Online].Z-alizadeh sani data set. Retrieved fromhttps://archive.ics.uci.edu/ml/datasets/Z-Alizadeh+Sani

[3] Alom, Z.; Azim, M.A.; Aung, Z.; Khushi, M.; Car, J.; Moni, M.A. Early Stage Detection of Heart Failure Using Machine Learning Techniques. In Proceedings of the International Conference on Big Data, IoT, and Machine Learning, Cox’s Bazar, Bangladesh, 23–25 September 2021.

[4] Arunpradeep, N.; Niranjana, G. Different Machine Learning Models Based Heart Disease Prediction. Int. J. Recent Technol. Eng. IJRTE 2020, 8, 544–548, ISSN 2277-3878.
[5] Available online: https://www.potentiaco.com/what-is-machine-learning-definition-types-applications-and-examples/ (accessed on 10 January 2022).
[6] Framingham Heart study dataset. https://kaggle.com/amanajmera1/ framinghamheart-study-dataset. Accessed Jan. 20, 2023.
[7] Golande, A.; Pavan Kumar, T. Heart disease prediction using effective machine learning techniques. Int. J. Recent Technol. Eng. IJRTE 2019, 8, 944–950, ISSN 2277-3878.
[8] Gour, S.; Panwar, P.; Dwivedi, D.; Mali, C. A Machine Learning Approach for Heart Attack Prediction. In Intelligent Sustainable Systems; Springer: Singapore, 2022; pp. 741–747.
[9] Johnson, K.W.; Torres Soto, J.; Glicksberg, B.S.; Shameer, K.; Miotto, R.; Ali, M.; Ashley, E.; Dudley, J.T. Artificial intelligence in cardiology. J. Am. Coll. Cardiol. 2018, 71, 2668–2679.
[10] Juhola, M.; Joutsijoki, H.; Penttinen, K.; Shah, D.; Pölönen, R.P.; -Setälä, K. Data analytics for cardiac diseases. Comput. Biol. Med. 2022, 142, 105218.
[11] Kataria R, Meena SK. Machine learning techniques for heart disease prediction: a comparative study and analysis. Health Technol 2021;11:87–97.
[12] Maini E, Venkateswarlu B, Gupta A. Applying machine learning algorithms to develop a universal cardiovascular disease prediction system. Int.Conf.Intell. Data Commun. Technol.Internet Things 2018:627–32.
[13] Marimuthu, M.; Abinaya, M.; Hariesh, K.; Madhankumar, K.; Pavithra, V. A Aalto Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach. Int. J. Comput. Appl. 2018, 181, 975–8887.
[14] Modepalli, Kavitha, Gnaneswar G, Dinesh R, Sai Y, Suraj R. Heart Dis.Pred using Hybrid.Mach. Learn.Model. 2021:1329–33. https://doi.org/10.1109/
[15] Shah D, Patel S, Bharti SK. Heart disease prediction using machine learning techniques. SN COMPUT. SCI. 2020;1:345. https://doi.org/10.1007/s42979-020.
[16] Sharma, S.; Parmar, M. Heart diseases prediction using deep learning neural network model. Int. J. Innov. Technol. Explor. Eng. IJITEE 2020, 9, 124–137, ISSN 2278-3075.
[17] Tuncer, T., Dogan, S., Plawiak, P., &Subasi, A. (2022).A novel discrete wavelet-concatenated mesh tree and ternary chess pattern based ecg signal recognition method.Biomedical Signal Processing and Control, 72, 103331. doi:10.1016/j.bspc.2021.103331.
[18] UCI machine learning repository: heart disease data set. http://archive.ics.uci.edu /ml/datasets/Heart+Disease. accessed Apr. 15, 2023.
[19] Velusamy, D., &Ramasamy, K. (2021).Ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset.Computer Methods and Programs in Biomedicine, 198, 105770. doi:10.1016/j.cmpb.2020.105770
[20] World Health Organization. Cardiovascular Diseases (CVDs). Available online: https://www.afro.who.int/health-topics/ cardiovascular-diseases (accessed on 10 January 2022).
[21] Yadav, D. C., &Pal, S. (2020).Prediction of heart disease using feature selection and random forest ensemble method.International Journal of Pharmaceutical Research, 12(4), 56–66.