International Journal of Data Warehousing and Mining
Volume 18 • Issue 1

An Ensemble Approach for Prediction of Cardiovascular Disease Using Meta Classifier Boosting Algorithms

Sibo Prasad Patro, GIET University, India*
Neelamadhab Padhy, GIET University, India
https://orcid.org/0000-0002-8512-3469
Rahul Deo Sah, Dr. Shyama Prasad Mukherjee University, India

ABSTRACT

There are very few studies are carried for investigating the potential of hybrid ensemble machine learning techniques for building a model for the detection and prediction of heart disease in the human body. In this research, the authors deal with a classification problem that is a hybridization of fusion-based ensemble model with machine learning approaches, which produces a more trustworthy ensemble than the original ensemble model and outperforms previous heart disease prediction models. The proposed model is evaluated on the Cleveland heart disease dataset using six boosting techniques named XGBoost, AdaBoost, Gradient Boosting, LightGBM, CatBoost, and Histogram-Based Gradient Boosting. Hybridization produces superior results under consideration of classification algorithms. The remarkable accuracies of 96.51% for training and 93.37% for testing have been achieved by the Meta-XGBoost classifier.

KEYWORDS

AdaBoost, CatBoost, Classification Algorithms, Fusion-Based Ensemble Model, Gradient Boosting, Histogram-Based Gradient Boosting, Hybridization, LightGBM, XGBoost

1. INTRODUCTION

In the last few decades, cardiovascular disease has become the main cause of death in the world. According to the WHO (World Health Organization), 17.9 million people died worldwide in 2019 from cardiovascular disease, accounting for 32% of all deaths. 85 percent of these deaths are caused by heart attacks and strokes. Since the last three-quarters of cardiovascular deaths took place in low and middle-income countries. 17 million premature deaths are due to no communicable diseases in the year 2019 and 38% of deaths are caused by CVDs. Cardiovascular diseases are high morbidity, high mortality, and high disability in nature. According to the European Society and Cardiology department, in a year nearly 3.6 million people are being diagnosed around the world (Coats, A. J. S., 2019; Spoletini, I., & Seferovic, P., 2017) Most affected by heart disease are from United States (US) (Heidenreich, P. A., 2011). Even with advanced techniques and perfect treatment, 50% of patients cannot fully care for themselves. The popularity of cardiovascular disease becomes a global problem today. Breath shortness, swollen feet, physical body weakness are the common symptoms of heart

DOI: 10.4018/IJDWM.316145 *Corresponding Author

disease (Durairaj, M., & Ramasamy, N., 2016). Many studies are being conducted to predict heart disease early using various machine learning algorithms and approaches. But the existing techniques are not much reliable in the context of execution time and accuracy. Due to the deficiency of technology and medical experts, heart disease diagnosis and treatment become more difficult. An effective and accurate diagnostic technology can save a large number of patient's life (Al-Shayea, Q. K., 2011). The chances of heart disease patients survive is a maximum of 1-2 years. Generally, the physician checks the patient's symptoms, physical examination reports, and symptoms for heart disease diagnosis. The outcome of these traditional techniques is not much effective and accurate to detect heart disease in a patient body. However, these approaches are more expensive (Tsanas, A., 2011).

In the medical domain, data mining and artificial intelligence (AI) techniques have been widely used to solve problems in urology, cancer, liver pathology, gynaecology, liver pathology, and perinatology (Sarwar, A., 2015). AI aims to build an AI-based system, that can assist a physician with an expert diagnosis as well as helps to increase the prediction probability of the disease in the patient body. AI algorithms play a vital role in exploring various diseases in the hidden pattern of the dataset. Using a data mining model such patterns are ready for clinical diagnosis (Soni, J., 2011). This prospective technique helped to build an expert system that can be used to predict and detect cardiovascular disease in the patient body. Recent research has used machine learning techniques to diagnose and predict various cardiac problems. Melillo et al. (Melillo, P., 2013) made a contribution to an automatic classifier for patients with congestive heart failure (CHF) that distinguishes between those at low and high risk. The sensitivity and specificity of the classification and regression tree (CART) were 93.3 percent and 63.5 percent, respectively. Rahhal et al. (Al Rahhal 2016) suggested a deep neural network (DNN) classification of ECG signals to learn the feature subset and improve performance. Guidi et al. (Guidi, G., 2014) made a significant contribution to a clinical decision support system (CDSS) for heart failure analysis (HF). They evaluated by comparing the performance of various machine learning classifiers, such as neural network (NN), support vector machine (SVM), a system with fuzzy rules that uses CART, and random forests (RF). The CART model and RF performed the best, with an accuracy of 87.6 percent. Zhang et al. (Zhang, R., 2017) discovered a NYHA class for HF from unorganised patient records using natural language processing (NLP) and the rule-based method, with a 93.37 percent accuracy. The high dimensionality of the dataset is a major issue in machine learning (Pedregosa, F., 2011). Because the analysis of many features requires a huge amount of memory and leads to overfitting, the weighting features reduce redundant data and processing time, thus improving the algorithm's performance. Dimensionality reduction transforms and simplifies data using feature extraction, whereas feature extraction minimises the dataset by removing the unnecessary features (Lahner, E., 2008).

In recent years one of these concepts is come up with the plan of integrating multiple classifiers as a new concept for improving the performance of individual classifier (Mohapatra, S., 2014). These classifiers are based on various classification approaches and they are able to accurately classify samples at varying rates. Such kind of classifiers is termed ensemble classifiers. They have the potential to increase the performance of several base-level classifiers or weak classifiers. The ensemble classifiers help to train them to improve the performance level of the base or weak classifiers (Ardakany, A. R., 2010, December) . In recent years, a hybrid ensemble classifier technique is designed, which is better than an ensemble classifier. Hybrid ensemble classifier proved to achieve better performance than the existing basic ensemble algorithms (Woźniak, M., 2014). Ensemble learning algorithms include bagging, boosting, and random forests. They all have one thing in common: they generate ensembles of base classifiers and ensure their diversity by feeding them different sets of learning examples.

For an accurate diagnosis of cardiac disease, the development of a machine learning-based clinical diagnostic system is necessary (Gonsalves, 2019). An expert decision system based on machine learning and Artificial Fuzzy Logic (AFL) applications can accurately detect heart disease patients, potentially lowering the death rate (Ansarullah, S. I., 2019). Many researchers utilise the Cleveland heart disease data set to forecast heart disease. There should be correct data required by

the prognostic machine learning model for training and testing. The goal of creating a data collection for training and testing is to increase the machine learning model's performance. Even identifying proper features from the data set helps to improve the performance of the model and prediction results. Feature selection and data balancing are some of the most important key parameters to enhance the performance of the model. Many machine learning-based heart disease detection diagnostic techniques were proposed using several machine learning algorithms, but these techniques having some loopholes. These techniques are failed in training large data sets, inconsistency inaccuracy, balancing the data, and many more. Hence they are not much effective for diagnosing heart disease. Data standardization helps for improving the prognostic capabilities of machine learning models. The Min-Max Scalar data preparation technique removes missing features from the dataset, whereas the Standard Scalar technique improves the model's performance (Nahar, J., 2013). The Greedy Algorithm (GA), Principal Component Analysis (PCA), Local Learning-Based Features Selection (LLBFS), and other techniques are used to choose features. Feature selection is used to train the machine learning model faster by selecting important parameters. Additionally, there are various optimization techniques used in the models for the optimization of features (Li, J. P., 2020).

XGBoost algorithm was proposed by Chen. T in the year 2016 (Chen, T., & Guestrin, C. 2016). It has more attention in various domains. Zheng H. proposed the XGBoost technique for obtaining the characteristic in the area of the power system (Zheng, H., 2017). Torbay L. used the XGBoost method to analyze the neurophysiologic properties of the five different linguistic regions of the human brain for classifying the epilepsy type of patients. This research proved that the XGBoost algorithm obtains excellent performance. Many studies are with the combination of XGBoost with other methods to solve a problem. Several algorithms are designed for heart disease prediction. According to Soni et al. (Soni, J., 2011) among Neural Network, Naïve Bayes, and Decision tree, the decision tree is one of the best technique for heart disease prediction. However, the decision tree approach is difficult to use when dealing with continuous variables. As a result, an effective Gradient Boosting Decision Tree (GBDT) algorithm was created to handle with a variety of data types, including both continuous and discrete values (Zhao, L., 2018).

Many researchers proposed various classifiers for predicting cardiovascular heart disease by taking individual classifiers or meta classifiers. Individual classifiers may not yield satisfactory results; therefore, a meta classifier or ensemble classifier can be constructed to outperform individual classifiers. A meta classifier is a tool that aids in the training of many classifiers in order to anticipate the final expected outcome. Ensemble classifiers assist several classifiers in predicting disease in a reliable and sufficient manner. The combination of classifiers can be homogeneous or heterogeneous. Compare to the individual classifier, the meta classifier's performance is better (Woźniak, 2014).

In this paper, a novel hybridization approach known as the linear stacking model is designed. The model is the combination of different ensemble-based boosting algorithms. XGBoost, Ada Boost, Gradient Boosting, LightGBM, Cat Boost, and Histogram-Based Gradient boosting hybrid classifiers are used. The main objective of this research is to enhance prediction accuracy. A cardiac dataset from the UCI data repository is collected for the proposed research. The performance and results of our proposed model shows better than any existing models.

The main contribution of this paper is to design an effective method to predict as accurately as possible for heart disease, In particular, is called Coronary Heart Disease. Required steps are as follows:

I. Due to the dataset containing some null values, we have applied the mean method for data cleaning techniques to the dataset to enhance the accuracy of the proposed model.
II. The classification model is being fine-tuned to deliver reliable results.
III. The feature scaling technique is applied to improve the performance of the machine learning algorithm.
IV. Linear Discriminant Analysis feature selection technique is used in this work for reducing the number of features and Principal Component Analysis (PCA) is used for minimizing the

dimensionality of the dataset and increasing the interpretability. It also helps to minimize information loss.

V. Additionally, different hybrid approaches are applied to improve the testing rate and reduce the execution time.

VI. The performances of the different models are evaluated based on the overall results with all meta-learning classifier boosting algorithms to enhance the prediction accuracy.

The flow of the paper is as follows: Literature review in Section 2, background work of this paper with a detailed description of the algorithms is elaborated in section 3, details of our proposed system in section 4, finally the Results and discussion with some recommendations for future works presented in section 5.

2. LITERATURE REVIEW

Hybrid and ensemble machine learning techniques were proposed by many researchers in the past. In the following, we have briefly discussed various works in the area of heart disease prediction and its diagnosis.

T.A. Asfaw et al. (Asfaw, T. 2019) highlighted comparing the performances of various classification algorithms for heart disease prediction. From a heart disease data set 10 attributes are used for this research work. K-Nearest Neighbors and Gaussian Naïve Bayes classification algorithms were used on these attributes. The mean and standard deviation of each attribute is calculated at the training stage. The proposed model produced an accuracy of 77.05% using the Gaussian Naïve Bayes classification algorithm.

L. Yang et al., (Yang, L., 2020) highlighted Cardio Vascular Disease (CVD) prediction is one the most effective measure for CVD control. In this research, a Random forest-based heart disease prediction model was proposed. The outcome of the proposed model is substantial enhancements over the multivariate regression model. The data was collected from more than 100,000 residents in Zhejiang province using a questionnaire survey, physical examination. The model was compared with a multivariate regression model, regression tree, and Ada Boost. Finally, the multivariate regression model produced a benchmark performance result of 71.43%.

E.K. Hashi et al., (Hashi, E. K., 2020) suggested an approach that uses Logistic regression, Decision Tree, and Random Forest classification models to predict heart disease. The suggested model applies a grid search method to all of the given classification algorithms to assist tune the hyperparameters. For this research work, the Cleveland Heart Disease dataset was collected from the UCI machine learning repository. The heart disease dataset is used for both training and testing purposes. After testing the dataset individual samples are classified. The traditional system is performed without any hyperparameter tuning method but the proposed system is performed with the help of grid search. In the cross-validation approaches the hyperparameters are optimized and tuned. Without hyper parameters tuning, the LR, KNN, SVM, DT, and RF classifiers produced an accuracy rate of 88.52%, 90.16%, 88.52%, 81.97%, and 85.25% respectively and with the hyperparameters tuning approach, the LR, KNN, SVM, DT, and RF classifiers in the refined set takes the accuracy rate 90.16%, 91.80%.

K.M. Almustafa et al., (Almustafa, K. M. 2020) carried out a comparative analysis of various classifiers for the classification of heart disease datasets to classify and predicting heart disease with minimal attributes. For this research, the dataset is collected from Cleveland, Hungary, Switzerland, and Long Beach. The dataset contains 1025 patients data with 76 attributes. Out of these only 14 attributes were used. The proposed system was evaluated with Naive Bayes, Decision tree J48, K-Nearest Neighbor (K-NN), JRip, SVM, Adaboost algorithms. To show the performance of the selected classification algorithm Stochastic Gradient Decent (SGD) and Decision Table (DT) classifiers are used. The performance of this model produced a result of 93.85% for the Decision Tree classifier.

H. Kahramanli et al. (Kahramanli, H., & Allahverdi, N. 2008) introduced a new hybrid technique that includes a fuzzy and artificial neural network and uses a hybridize neural network and for this proposed work the data were obtained from the University of California at Irvine (UCI) machine

learning repository. The obtained dataset contains a total of 303 samples of patients with heart disease. The proposed method performance was evaluated by medical classification studies. K-fold cross-validation technique applied to the dataset for obtaining the classification accuracies. The proposed method produced an accuracy of 82.45%.

R. Das et al. (Das, R., 2009) suggested a hybrid technique based on a neural network ensemble, which integrates neural network and ensemble-based methods. For this research work, the Cleveland dataset is used that composed of 14 columns and 297 rows. In this work, they introduced a methodology that uses SAS-based software 9.1.3 for diagnosing heart disease. For data preprocessing, they used SAS enterprise guide 4.3 and SAS enterprise miner 5.2 programs for analyzing heart disease by combining several neural networks with various ensemble nodes. Using ensemble model three independent neural network models were constructed. For heart disease diagnosis, the test results obtained 89.01% classification accuracy, 80.95% sensitivity, and 95.91% specificity.

Heart diseases are one of the most common problems globally, according to R. El Bialy et al., (El Bialy, R., 2016) and hence it requires a high level of accurate analysis and prediction. A new framework was proposed with the best ensemble combination method for diagnosing heart diseases using bagging, boosting, and staking. Bayes Net, Multilayer perceptron, Naïve Bayes, Decision tree classifiers, and Sequential Minimal Optimization classification algorithms were used using ensemble techniques for this research. On two benchmark data sets collected from two separate resources, the findings are examined using the Weka tool. The classification accuracy outperforms more than 90% in some cases.

Karthikeyan, G. et al., (Karthikeyan, G., 2021) said prediction of cardiovascular disease is an extremely challenging factor in the clinical decision support system(CDSS). The researcher proposed a Hybrid Linear stacking model for feature selection and XGBoost algorithms for the classification of heart disease. The model name was given as HLS-XGBoost. The model predicts the most influencing features of heart disease and is classified using the XGBoost algorithm for increasing the performance of the prediction accuracy. For this research, they collected the data from the UCI machine learning repository. In the dataset due to 6 records are missing out of 303 records, those records are removed. The proposed model HLS-XGBoost is compared with traditional approaches like LR, NB, MLP, DT, SVM, RF, and HDPM models respectively. The model is initiated with various feature subset combinations and diverse classification approaches. This model produces an improved performance level with 96%accuracy using the HLS-XGBoost model.

By merging numerous classifiers, Latha, C. B. C., et al., (Latha, C. B. C., 2019) suggested an ensemble classification method for enhancing the accuracy of weak algorithms. For the experimental work, the Cleveland heart dataset was identified from the UCI machine learning repository. The dataset consists of 14 attributes and 303 instances. An ensemble strategy involving bagging, boosting, voting, and stacking algorithms is utilized to improve prediction accuracy in heart disease. Bagging algorithms include Nave Bayes, Random Forest, Bayes Net, C4.5, multilayer perceptron, and PART. Adaboost and M1 algorithms are used for boosting. The outcome of the model was improved by a maximum of 6.92% using bagging and 5.94% using boosting. The Nave Bayes classifier has the highest accuracy of 83.17%, while C 4.5, Multilevel Perceptron, and PART have lower accuracy of less than 80%.

S Mohan et al. (Mohan, S., 2019) introduced a technique called Hybrid Random Forest with Linear Model (HRFLM) for enhancing the accuracy of heart disease prediction by using machine learning techniques to uncover key features. The predictive model is designed with several classification techniques. The researcher initiated hybrid RF with a linear model to improve the model functionality. The dataset was taken from the Cleveland UCI repository. To perform heart disease classification they used R studio rattle software. The proposed HRFLM produced an accuracy of 88.7% for heart disease prediction.

Budholiya, K et al.,(Budholiya, K., 2020) presented a diagnostic system model using an optimized XGBoost classifier for heart disease prediction. The researcher employed the Bayesian optimization technique to optimize the hyper-parameter of XGBoost, and the One-Hot (OH) encoding technique to encode the categorical features of the dataset to improve the performance of heart disease prediction. The model is tested using the Random Forest(RF) and Extra Tree(ET) classifiers on the Cleveland

heart disease dataset. Accuracy, specificity, sensitivity, F1-score, and AUC are some of the evaluation measures used for performance evaluation. They compared the proposed model to existing tree-based ensemble machine learning approaches, finding that the present model surpasses the prior work by 3.28 percent. The proposed model can be compared with other datasets to evaluate and produce better accuracy.

A cloud-based intelligent system powered by the SVM model was proposed by M.A. Khan et al.,(Almustafa, K. M. 2020).Cloud computing, Support Vector Machine (SVM), and some machine learning algorithms were all used in the suggested model. Electronic health records are used to manage the high volume of data using cloud computing and Machine learning techniques are used to extract hidden patterns and data analysis. Due to the volume of data is more diagnostic accuracy promptly with appropriate treatments are required. A cloud-based intelligent system empowered by the SVM model gave 93.33% accuracy.

Heart disease is one of the main causes of death worldwide, according to Nasarian, E. et al., (Nasarian, E., 2020). Heart disease is associated with high healthcare expenditure. Heart disease datasets contain various features with different degrees of association with heart disease. The decision tree (DT), Gaussian Naive Bayes (GNB), Random Forest (RF), and XGBoost classifiers were used to create a heterogeneous hybrid feature selection (2HFS) approach for feature extraction. The proposed model was evaluated on Nasarian CAD dataset. To deal with the unbalanced data The techniques of synthetic minority over-sampling (SMOTE) and adaptive synthetic (ADASYN) were applied. With SMOTE and the XGBoost classifier, the model achieved an accuracy of 81.23 percent in classification.

Tama, B. A., et al.,(Tama, B. A., 2020) proposed a Coronary Heart Disease (CHD) detection technique using machine learning ensemble classifier algorithms. With multiple classifiers, a two-tier ensemble technique was created, with some ensemble classifiers serving as base classifiers for another ensemble. Random forest, gradient boosting machine, and extreme gradient boosting classifier were used to create a stacked architecture. The proposed model was evaluated on Z-Alizadeh Sani, Statlog, Cleveland, and Hungarian datasets. For identifying the most significant feature of a dataset particle swarm optimization feature selection approach was used. The performance of the model was evaluated with an accuracy rate of 83.90%.

Many of the researchers carried out different heart disease prediction models using different machine learning classifiers. Few of them used ensemble and hybrid approaches, but the performance of the models was unsatisfactory. Due to the increase in the number of heart disease patients day by day, the size of the database increased drastically but the data are not maintained properly for analysis. In such a case is it very difficult for the researchers to obtain value from data. Many of the research works are conducted on a single dataset that contains a set of 14 attributes from the UCI repository which leads to increased costs due to more number of attributes. Many existing research works are carried out to build an efficient system for heart disease prediction but many key challenges have been left untouched and that may improve the quality of heart disease prediction to a great extent. Table 1 is a comparison table for the strategies discussed in related studies.

As per the above Table 1. we come to conclude that all the above-said techniques proved their efficiency in their approaches, but there are certain issues identified which can be resolved. In this approach, an extensive literature survey has been carried out and challenges are identified. Looking at the challenges a model has been proposed that focuses on working towards the challenges which are identified. The major objective of the work was to build an efficient heart disease prediction system that identifies the risk accurately. To achieve the objective, a model using a hybrid of UMAP and XGBoost has been proposed in this dissertation. The major contributions of this work are listed below.

The above-mentioned approaches, few of the work are not adopted with feature attribute selection and balancing data at the time of testing and training the data to increase the accuracy of the prediction model on heart disease datasets. We need to address the regularization overfitting and underfitting problem. The previous approaches and techniques addressed real improvement in the results; anyhow there are still some techniques and better performance in the results can be expected, such as:

Table 1. Comparison of diverse existing heart disease prediction models

Sl. No.	Author	Year	Model	Algorithms used	Performance matrix
1	T.A. Asfaw et al	2019	Heart disease prediction model	K-Nearest Neighbors, Gaussian Naïve Bayes classification	77.05%
2	L. Yang et al.,	2020	A Random forest-based heart disease prediction model	Random forest, regression tree, and Ada Boost	71.43%
3	E.K. Hashi et al.,	2020	Prediction of the heart disease model	Logistic regression, Decision tree, and Random Forest classification models	88.52%
4	K.M. Almustafa et al.,	2020	A classification and heart disease prediction model	Naive Bayes, Decision tree J48, K-Nearest Neighbor, JRip, SVM, and Adaboost algorithms	93.85%
5	Kahramanli, H., et al.,	2008	A new hybrid technique model for heart disease prediction	A Fuzzy and artificial neural network	82.45%
6	Das, R et al.,	2009	A hybrid model	Neural networks and ensemble techniques	89.01%
7	R. El Bialy et al.,	2016	An ensemble-based framework	Bagging, boosting, and staking. Bayes Net, Multilayer perceptron, Naïve Bayes, Decision tree classifiers	90.00%
8	Karthikeyan, G. et al.,	2021	A Hybrid Linear stacking model	HLS-XGBoost	96.00%
9	Latha, C. B. C., et al.,	2019	An ensemble classification model	Naïve Bayes, Random Forest, Bayes Net, C4.5, multilayer perceptron, and PART	83.17%
10	Mohan, S et al.,	2019	A Hybrid Random Forest with Linear Model	Hybrid RF	88.70%
11	Budholiya, K et al.,	2020	A diagnostic system model	XGBoost, Bayesian optimization, Random Forest	91.80%
12	M.A. Khan. et al.,	2020	A cloud-based intelligent system	SVM	93.33%
13	Nasarian, E. et al.,	2020	Heart disease prediction model	Decision tree (DT), Gaussian Naive Bayes (GNB), Random Forest (RF), and XGBoost classifiers	81.23%
14	Tama, B. A., et al.,	2020	A Coronary Heart Disease (CHD) detection model	Random forest, gradient boosting machine, and extreme gradient boosting classifier	83.90%

A. None of the above-mentioned approaches explored the novel machine learning algorithms like Extreme gradient boosting (XGBoost), AdaBoost, Gradient Boosting, The LightGBM, CatBoost, and Histogram Gradient Boosting.
B. None of the previous works used the categorical feature encoding technique to encode categorical features in the heart disease dataset
C. Till now, any of the previous work has not used the Bayesian optimization technique for optimizing the machine learning model; this approach is one of the most efficient search strategies.

The goal of this proposed research is to analyze the accuracy and error rate of the boosting algorithms to discover the best features.'

3. BACKGROUND

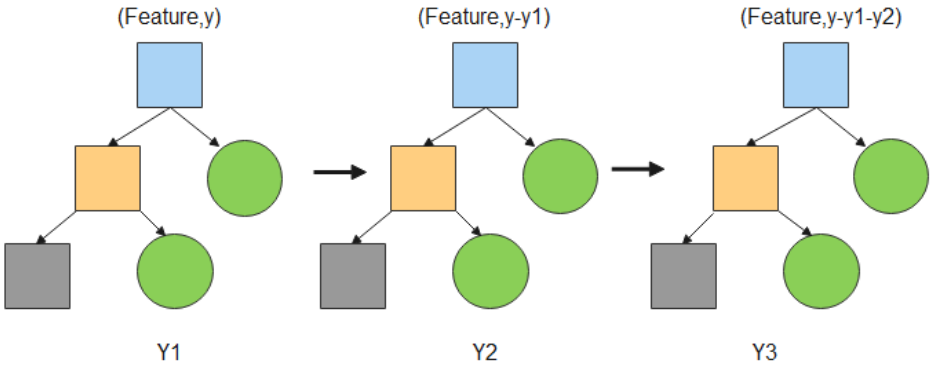
A theoretical framework for explaining machine learning termed a probably approximately correct (PAC) learning model (Valiant, L. G. 1984) includes boosting as one of the sources. Kearns and Valiant (Kearns, M. 1988; Kearns, M., & Valiant, L. 1994) were the researchers who initially thought about whether a "weak" learning algorithm performs better than random guessing in the PAC model can be "boosted" into an absolute accurate "strong" learning algorithm. The first polynomial-time boosting algorithm was proposed by Schapire et al., in 1989 (Schapire, R. E. 1990). After one year Freund (Freund, Y. 1995) developed more efficient boosting that gives the result in an optimal time. The first-ever experiment with boosting algorithms was implemented by Drucker, Schapire, and Simard (Drucker, H., 1993) on an OCR task. In our paper, we have used some novel boosting algorithms; XGBoost, Ada Boost, Gradient Boosting, LightGBM, Cat Boost, and Histogram-Based Gradient boosting hybrid classifier are used.

3.1 XGBoost

The gradient boosting approach is efficiently implemented with XGBoost. It is one of the most well-developed versions of Gradient boosting for improving accuracy. It contains a linear model as well as a tree learning algorithm. XGBoost and Gradient Boosting are ensemble tree methods. They are useful to boost weak learners. GB (Gradient Boosting) is a stage-wise additive modeling (Khare, P., 2013).

In this XGBoost technique, a weak classifier is first fitted to the data, and then a weaker classifier is fitted to improve the current model's performance without impacting or changing the prior classifier. This process continues till the end. In this continuous process, the new classifiers identify the performance of the previous classifier and need to identify where the previous classifiers are not performed well. The general flow of the Boosting algorithms is shown in Fig 1. It clearly indicates that y1 is computed by fitting the data to a decision tree, and then y-y1 is derived by fitting the second tree depending on the

Figure 1. The general architecture of XGBoost



findings of the previous phase. The same procedure is followed until the desired outcome is obtained.

The XGBoost objective function at iteration ***t*** that we need to minimize is shown in equation (5)

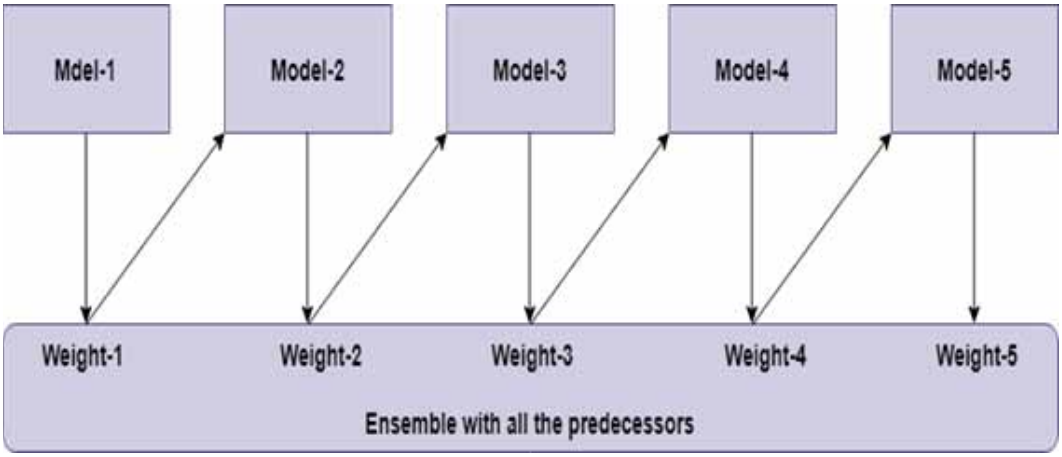
$$\mathcal{L}^{t} = \sum_{i=1}^{n} l\left(y_i, \widehat{y}_i^{(t-1)} + f_t\left(X_i\right)\right) \copyright \left(f_t\right) \tag{5}$$

Can be defined as $f\left(x + " x\right)$ where $x = \widehat{y}_i^{(t-1)}$

3.2 ADABoost

In machine learning, AdaBoost is applied as an ensemble approach. Adaptive Boosting is another name for it. The weights are reassigned to each instance in this method, with higher weights being allocated to erroneously classified instances. For supervised learning, boosting is used to lower the bias value as well as the variance. It is based on a method in which the learners are taught sequentially. Here each subsequent learner grew from previously grown learners except the first one. Finally, the weak learners are converted into strong ones. The working model of AdaBoost is shown in Fig. 11. It is a boosting technique that provides an approach for combining decisions, it is not necessarily a strong classifier to produce more accurate predictions (Freund, Y., 1999). An up-weighting or down-weighting process is used to alter the weights of numerous training data points iteratively. The algorithm generates a new base classifier in each iteration to best fit the current weighted samples and the weights are updated accordingly. This will help the misclassified samples to assign a higher weight to influence the training of the next base classifier. Finally, a weighted combination of the base classifier is the output of AdaBoost. Fig. 2 shows the working model of AdaBoost.

Figure 2. The working model of AdaBoost



The main goal of the algorithm is to maintain the distribution of the set of weights over the training set. The weights of the given distribution on training I on round t can be defined as Dt(i). In the beginning, all the weights are set equally, later in each round the weight of the incorrect classifier is increased, by this process the weak learners improve their efficiency in the training set. The job of the weak learners is to find a weak hypothesis $h_i : X \to \{-1, +1\}$ for an appropriate distribution D_i. The weak hypothesis is measured by its error shown in the following equation (5)

$$\in_t = Pr_{i\sim D_t}\left[h_t\left(x_i\right) \neq y_i\right] = \sum_{i:h_t\left(x_i\right)\neq y_i} D_t\left(i\right). \tag{5}$$

From the above equation, it is noticed that the error is measured with D_t. The error value occurred when the weak learner was trained. The weak learners are assigned weights D_t on the training phase. A subset of the training examples can be selected according D_t and the unweighted examples are used to train the weak learners.

3.3 Gradient Boosting

A weak learner can be converted into a strong learner using Gradient Boosting Algorithm (GBM). In this approach, each new tree fits on a modified version of the original data set. Each observation is assigned with equal weight at the initial stage of the algorithm. We can increase the weights of those observations that are difficult to classify once the first tree has been reviewed. Then the second tree is grown on those weighted data. Using this technique we can improve the prediction of the first tree and the new model which is generated that is the combination of $Tree1 + Tree2$. When we calculate the classification error from the new 2-tree ensemble model and develop a new tree to forecast the existing corrected findings, we get the new results. This process can be continued till the end of our model construction. Gradient descent based method is used to decide the step size or alpha, the alpha can be calculated at each iteration m, first pseudo residual(rim) is calculated and the new model hm(x) will construct from {xi, rim}. The pseudo residual calculation is shown in equation (6).

$$r_{im} = \frac{\partial L}{\partial + \left(x_i\right)}\left(y_i + \left(x_i\right)\right) \mid f^x = f_{m-1}\left(x\right) \tag{6}$$

Now the alpha can be calculated for minimizing the loss function, shown in equation (7).

$$\mathcal{L} = argmin_{\mathcal{L}} L(y_i + m_{-1}\left(x1 + \mathcal{L}h_m\left(x1\right)\right) \tag{7}$$

Gradient Boosting is a decision tree-based model, that can be defined as $f\left(x\right)$ and $h\left(x\right)$ are treated as cart trees, for constructing a tree with T leaves, model $hm\left(x\right)$ shown in the equation (8)

$$fm_x = f_{m-1}\left(x\right) + \sum_{j=1}^{t} \mathcal{L}_{jm} \mid R_{jm}^{(x)} \tag{8}$$

$$\mathcal{L}_{jm} = arg\,min_{\mathcal{L}} \sum_{x_i \epsilon R_{jm}} L\left(y_i, f_{m-1}\left(x_i\right) + \mathcal{L}\right) \tag{9}$$

3.4 Light Gradient Boosting

It's a variation on the gradient boosting method. It's a type of machine learning algorithm that may be used to solve classification and regression problems. It adds automatic feature selection and focuses on boosting examples with larger gradients. The decision tree model is used to create the ensembles. The freshly generated trees were introduced to the ensemble one by one and fitted to fix the previous models' prediction errors. This type of ensemble machine learning concept is called boosting. This procedure assists in the acceleration of training and the improvement of prediction performance. The models are fitted using any arbitrary differentiable loss function and the gradient descent optimization procedure in this technique.

The mathematical analysis for the GOSS technique is shown in equation (10)

$$\overline{v}_j(d) = \frac{1}{n}\left(\frac{(\sum_{x_i \epsilon A_l} g_i + \frac{1-a}{b}\sum_{x_i \epsilon B_l} g_i)^2}{n_i^j d} + \frac{(\sum_{x_i \epsilon A_r} g_i + \frac{1-a}{b}\sum_{x_i \epsilon B_r} g_i)^2}{n_i^j d}\right) \quad (10)$$

In the above expression,

$A_i = \{ x_i ? A : x_{ij} ? d\}$, $A_r = \{ x_i ? A : x_{ij} > d\}$, $B_i = \{x_i ? B : x_{ij} ?d\}$,

$B_r = \{x_i ? B : x_{ij} > d\}$, and the coefficient $\frac{(1-a)}{b}$ is used to normalize the sum of the gradients over B back to the size of A^c.

3.5 CATBoost

CatBoost algorithm used for solving problems such as regression, classification, ranking, and multi-class classification, etc. This algorithm is used as a gradient boosting technique on the decision tree. It is easy to use and works well with heterogeneous data and even relatively small data. It essentially creates a strong learner from an ensemble of many weak learners. CatBoost is a gradient boosting algorithm that works with categorical features and outperforms other gradient boosting methods. At every newly constructed trees and target-based statistics, it employs one-hot max-size encoding principles and applies the permutation concept for translating the class labels into numbers utilising Greedy methods. The steps of the algorithm are as follows:

a) It groups the dataset into a random order.
b) Converts the class labels into an integer number
c) Converts the categorical values into numeric values.

RFC combines the bagging technique, bootstrap feature selection, and feature randomness to help combine decision trees and generate new trees. The final tree will be determined by the RFC results obtained from each tree's vote. The RFC result with the most votes will be designated the final tree.

Mathematically, the target estimate of the i^{th} the categorical variable of the k^{th} element of D is shown in equation (11)

$$\hat{x}_k^i = \frac{\sum_{x_j \epsilon d_k} 1x_k^i = x_{k,}^j y_j + ap}{\sum_{x_j \epsilon d_k} 1x_k^i = x_{k,}^j y_j + a}; \text{if } D_k = \{x_j : \sigma(j) < \sigma(i)\}^{\text{where a>0}} \quad (11)$$

In the above equation, the indicator function $1x_k^i = x_{k,}^j$ takes the value 1 when the given i^{th} component of CatBoost's input vector x_j is equal to the i^{th} component of the input vector $x_{\underline{k}}$. Here we use k as in the k^{th} the element according to the order we put on d with the random permutation A, and i takes on the integer values 1 through -1.

3.6 Histogram-Based Gradient Boosting

It is an ensemble machine learning algorithm. This algorithm helps to speed up the construction of each decision tree. The trees that are added to the ensemble can be radically accelerated by differentiating the continuous input values into hundred unique values. Except this, it helps to build an efficient data structure to represent the input data. Boosting is used to convert inaccurate weak learners into a single accurate forecast. A restricted family of functions F is the combination of base

learners and a general boosting algorithm is the combination of sequence of function $\left\{f_t\right\}_{t=1}^{T}$ from F to minimize the certain observed loss. The final prediction is shown in equation (12).

$$F = \sum_{t=1}^{T} w_t f_{t,} \tag{12}$$

In the above said equation the $w_t \geq 0,\ t = 1.....T,$ are the weights and $f_{t\in}F,\ t = 1.....T.$ from a functional gradient descent viewpoint represented in statistics (Tama, B. A., 2020). The boosting is a reconstructed step-by-step optimization problem with various loss functions. As per the said expression, Gradient boosting needs more computing when the negative functional gradient is the response and identifying a specific model from the listed class of function at each boosting iteration to update the predictor. The expression is shown in equation (13).

$$U_i = -\frac{\partial L\left(y_{i,} f\left(x_i\right)\right)}{\partial f\left(x_i\right)} \Big|_{f_{(x_i)=f(x_i)}} \tag{13}$$

4. PROPOSED WORK

In this section, the system architecture is presented. Fig. 3 represents the overview of the system architecture. The core module of the proposed system consists of Data collection, Data Preprocessing, Scaling and Normalize the feature, Dimensionality reduction using LDA and PCA, train the data set in phase-I using novel machine learning algorithms, testing the dataset in phase-II for meta classifier then finally prediction of heart disease in the final step.

4.1 Data Collection

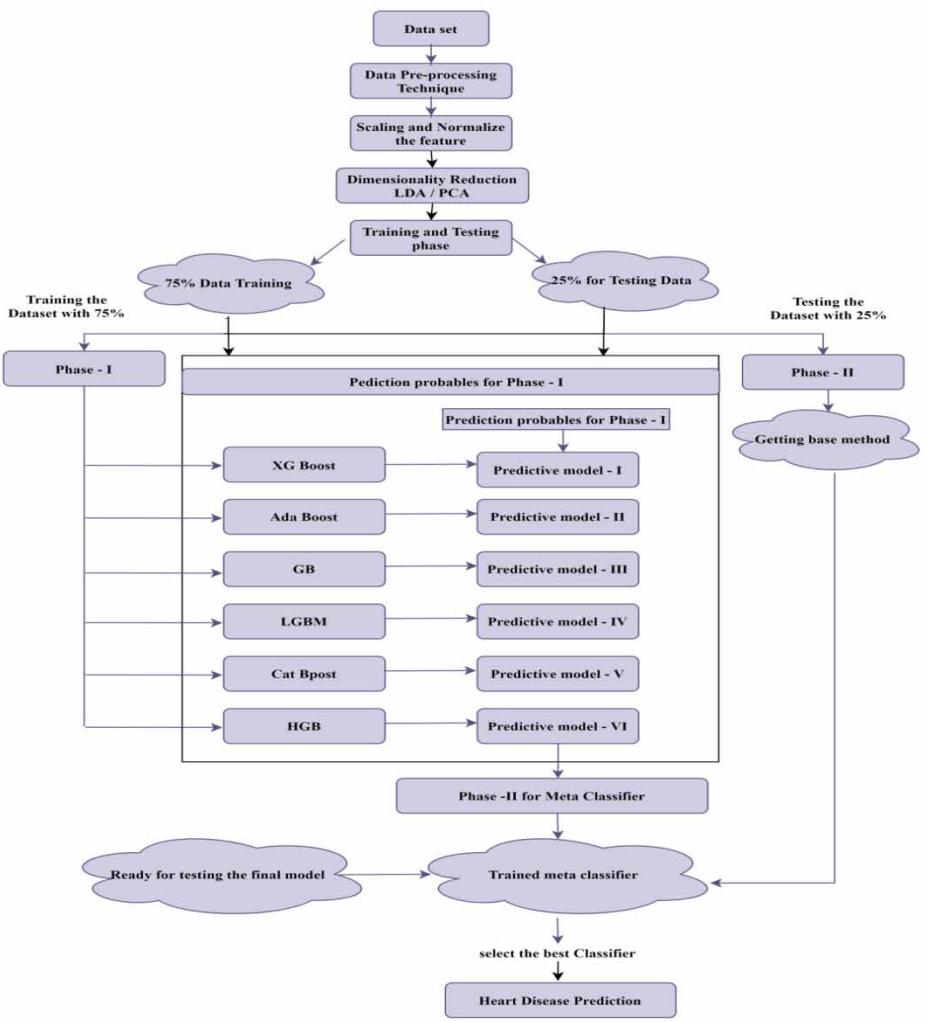
For this paper, from UCI Repository a dataset is taken. Formally, it is named as Heart Disease Dataset. Cleveland dataset is commonly used for many machine learning-based research works. The dataset contains 303 records with 14 medical features. The features and descriptions of each attribute are shown in Fig 4. In the given dataset the "Class" feature represents the presence or absence of heart disease. The original values of the features are initially 1,2,3 and 4 but later the values are transformed into one category i.e 1. When the value becomes 1 that represents the presence of heart disease.

Machine Learning Repository (UCI Databases) Heart disease dataset found at https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/heart-disease.names.

4.2 Data Preprocessing

For making the diagnosis of a disease quicker and easier, the database should be free from redundant, missing, and irrelevant data. In our research, after collecting the database, we applied to preprocess process for avoiding such as redundant, missing, and irrelevant data. The original dataset initially contains 303 patient records, where 7 records are with some missing or null values. The null values need to update with non-null values because the null values may lead to the wrong prediction for any machine learning model. Hence, we applied the mean method to replace the null values. The mean method is formulated as:

Figure 3. Heart disease prediction system architecture



$$Q(x) = \begin{cases} mean(x), & if\ x = null\ /\ missed \\ x, & otherwise \end{cases} \quad (1)$$

The x in the above equation (1) represents the instances of feature vectors that lie in n-dimensional space, $x \in R$.

The training set is made up of 75% of the data, whereas the test set is made up of 25% of the data. Ten-fold cross-validation is used to test the model's generalisation capacity. For normalizing the features two different maximum and minimum methods are used (current-min) (max-min).

4.3 Scaling And Normalizing the Feature

One of the most important aspects in selecting the appropriate feature from a dataset is feature selection. This process involves selecting relevant features of original features based on predefined criteria. Filter, wrapper, and hybrid models are among the feature gathering algorithms used with

Figure 4. Detailed information of heart disease dataset

Num.	Code	Feature	Type	Description
1	Age	Age	Continuous	Age in years
2	Sex	Sex	Discrete	sex (1 = male; 0 = female)
3.	Cp	Chest pain type	Discrete	1 = typical angina; 2 = atypical angina; 3 = non-angina pain; 4 = asymptomatic
4	Trestbps	Resting blood pressure (mg)	Continuous	At the time of admission to the hospital [94, 200]
5	Chol	Serum cholesterol (mg/dl)	Continuous	Multiple values between [Minimum Chol: 126, Maximum Chol: 564]
6	FBS	Fasting bood sugar > 120 mg/dl	Discrete	1 = yes; 0 = no
7	Restecg	Resting electrocardiographic results	Discrete	0 = normal; 1 = ST-T wave abnormal; 2 = left ventricular hypertrophy
8	Thalach	Maximum heart rate achieved	Continuous	Maximum heart rate achieved [71, 202]
9	Exang	Exercise induced angina	Discrete	1 = yes; 0 = no;
10	Oldpeak	ST depression induced by exercise relative to rest	Continuous	Multiple real number values between 0 and 6.2.
11	Slope	The slope of the peak	Discrete	1 = upsloping;

various assessment criteria (Ray, P., 2021). Many datasets contain multiple features and that cover different degrees of magnitude, units, and range. Different degrees of characteristics are particularly sensitive to machine learning techniques. As a result, the feature scalling technique is used to increase the machine learning algorithms' performance.

Normalization is a rescaling technique that shifts and rescales the data. As a result, the rescaled values fall between 0 and 1. This technique is called as Min-Max scaling. The following equation (2) representing the formula for normalization.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{2}$$

Xmax and Xmin are the maximum and minimum values of the feature, respectively, in the preceding equation..

A. When the X value is minimum in the column then X' is 0
B. When the X value is maximum in the column then X' is 1
C. When the X value lies between maximum and minimum then the value of X' lies between 0 and 1.

We can reduce the noise in a dataset by using Moving Average (MA). A sequence of averages of different subsets of the given dataset is generated using this technique. The moving average is calculated using the arithmetic means of the provided collection of variables. Equation (3) shows how to calculate the Moving Average:

$$MA = \frac{x1 + x2 + x3 + x4 + \ldots + xn}{N} \tag{3}$$

In the above equation, the $\mathrm{x1} + \mathrm{x2} + \mathrm{x3} + \mathrm{x4} + \ldots + \mathrm{xn}$ are instances of the feature vector and N is the total number of attributes. In normalization, the Z-score normalization technique is used to rescale the values of the attributes of the given dataset. With this technique, we normalize the distribution of data with zero means. It also helps to reduce the skewness of data distribution. The standardization formula is shown in equation (4)

$$\mathrm{R}\left(\mathrm{x}\right) = \frac{\mathrm{X} - \overline{\mathrm{X}}}{\tilde{\mathrm{A}}} \tag{4}$$

In the given equation the x is the instances of feature vectors with n-dimensional space, $\mathrm{x} \ \epsilon \ \mathrm{R}^{\mathrm{n}}$. $\overline{\mathrm{x}}$ represents mean and $\dot{\mathrm{A}}$ represents the standard deviation of the attributes.

4.4 Dimensionality Reduction Using LDA And PCA

By generating a collection of principal variables, the dimensionality reduction technique is used to reduce the number of random variables. When a dataset contains a large number of features, it may cause the learning model to become overfit. Various strategies are used to identify the required features and minimise the dataset's dimension. This process is divided into two approaches, they are feature selection and feature extraction. We employed Principle Component Analysis (PCA) and Linear Discriminant Analysis (LDA) as dimensionality reduction strategies in our study. PCA is a linear approach for dimensionality reduction. This technique works as a linear mapping of the data from a higher-dimensional space to a lower-dimensional space and the variance of the data in the low-dimension is maximized. LDA is a method to find a linear combination of features for separating the classes. It helps to identify the difference between the classes of data. This technique works on independent variables of each observation are in continuous quantities. LDA and PCA feature

selection algorithms are used to identify the important features from the dataset to make the system able to predict heart disease correctly. The framework of heart disease prediction is shown in Fig. 5.

A heatmap is used to analyze the data for this research. A comparison was applied to find the correlation values of the features. All the correlation values lie between 0.0 to 0.5. Therefore, we say the features have a good correlation.

The basic idea to use PCA is to remove the noisy and correlated variables to preserve the significance. The following Fig. 6 and Fig. 7 represent how many PCA are required. For our research work, we have taken two parts of results in a specific image representation. More than two components are not required. This performance of this process gives a minimum number of attributes with less error rate and better results.

The following Fig. 8 and Fig. 9 show the representation of PCA components. In Table 2 and Table 3 the PCA Accuracy and PCA accuracy wall time are represented respectively.

In Table 2, C stands for components, and 12 principal components exist. Out of the above 12 components, we can select the best component values to represent the whole data. The PCA accuracy-test wall time is shown in Table 3.

Linear Discriminant analysis is a statistical is a statistical-based pattern classification technique under unsupervised learning. LDA is useful in data preprocessing. This technique is used to minimize the attributes or features for the given dataset before the pattern classification. The relationship between chest pain and thalach is shown in Fig. 10.

In above Fig. 11, the X-axis represents chest pain (cp), and the Y-axis representing thalach. The figure represents the relationship between chest pain and thalach.

In the above Fig. 8, the X-Axis represents the cholesterol, and Y-axis represents a person's age. We can understand that people in the age group of 30 have more cholesterol, leading to chest pain. This causes to heart attack and may lead to death. The above figure is also representing a person who has less cholesterol with more age. The LDA accuracy-test wall time is shown in the following Table 4.

Figure 5. The framework of heart disease prediction

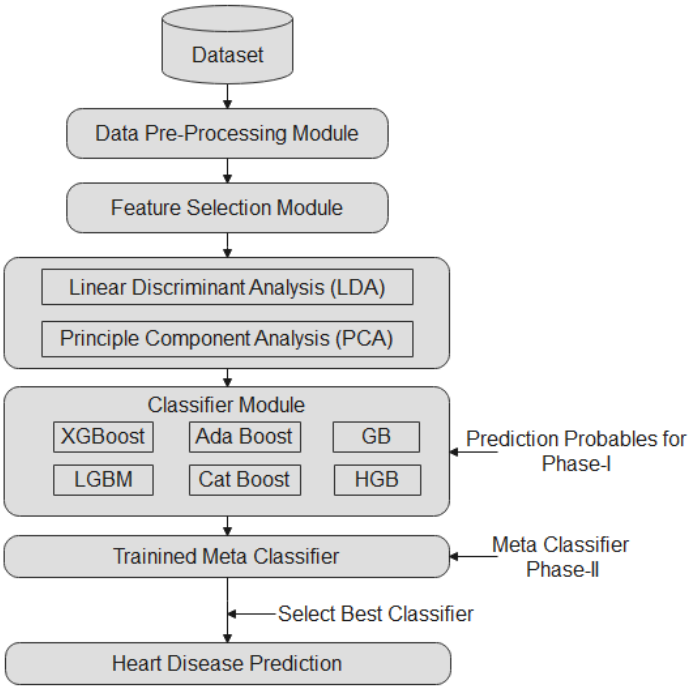


Figure 6. PCA Componet1

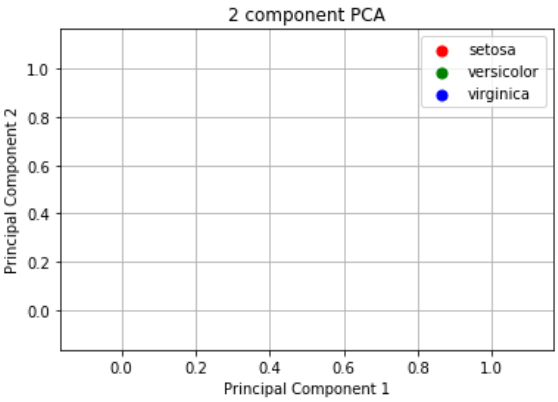


Figure 7. PCA Componet2

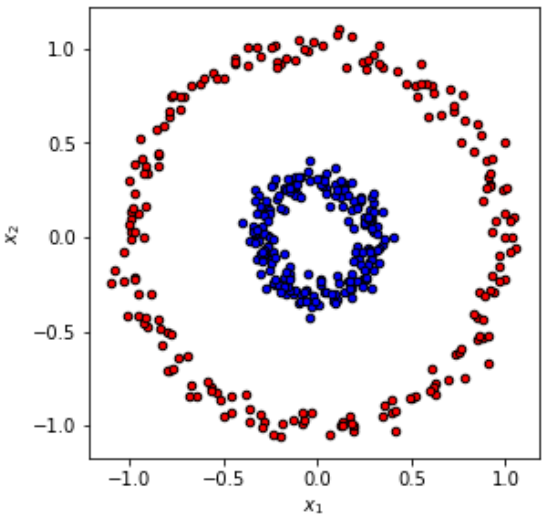


Figure 8. Projection by KPCA

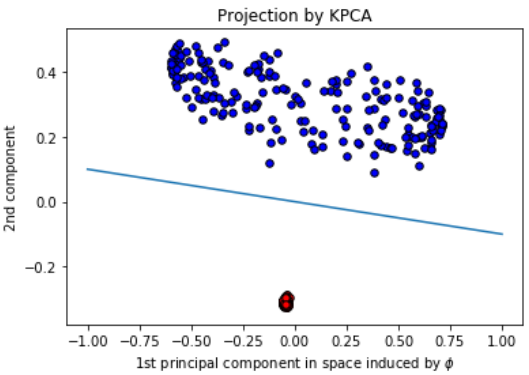


Figure 9. Project by KPCA

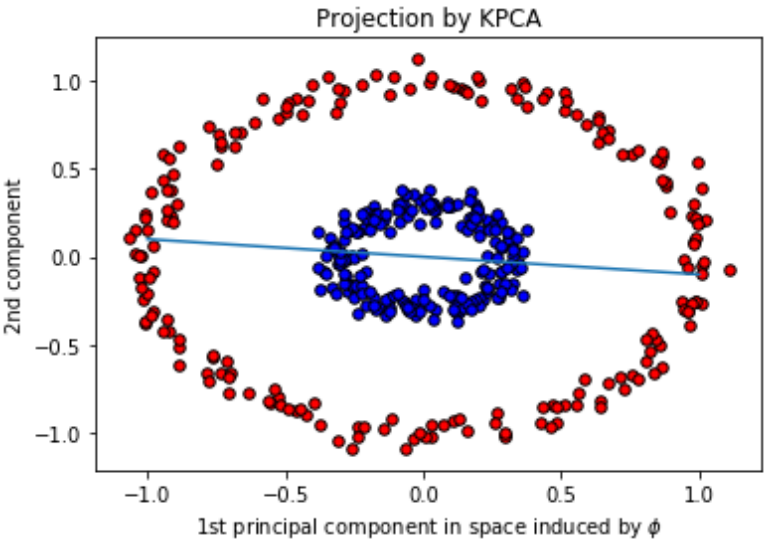


Table 2. PCA accuracy

Principal components (PCs)	*Accuracy*
C1:	0.737705
C2:	0.770492
C3:	0.836066
C4:	0.852459
C5:	0.852459
C6:	0.836066
C7:	0.819672
C8:	0.836066
C9:	0.852459
C10:	0.836066
C11:	0.836066
C12:	0.819672

Table 3. PCA Accuracy test wall time

Test set accuracy	**Wall time**
0.7704918032786885	444 ms
0.8524590163934426	466 ms

Figure 10. Identifying the relationship between chest pain and thalach

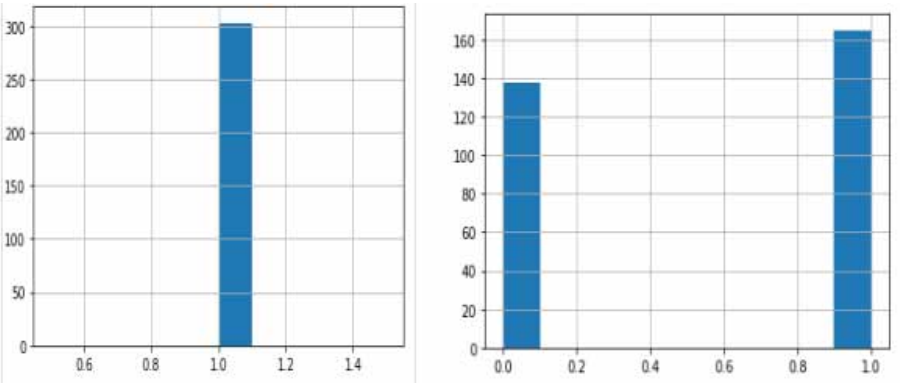
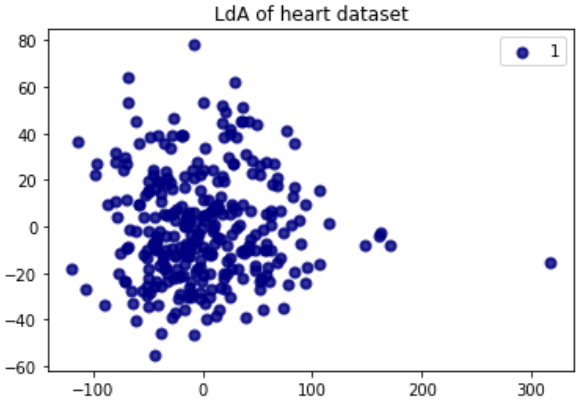


Figure 11. For LDA heart data set



4.5 Training and Testing

For this paper, the Heart Disease Dataset is split into two parts using the data splitting technique. One part is the training set with 75% for train the model and the second one is the testing set with 25% for evaluating the model. The training set is applied with 10 fold cross-validation technique. For training the model we have taken novel boosting machine learning algorithms. The proposed hybrid-ensemble built upon various classifier ensembles, i.e XGBoost, AdaBoost, Gradient Boosting, Light GBM, CatBoost, and Histogram-based Gradient Boosting. The boosting algorithms are used in prediction probable phase-I and Meta classifier are used in phase-II to select the best Classifier for heart disease prediction.

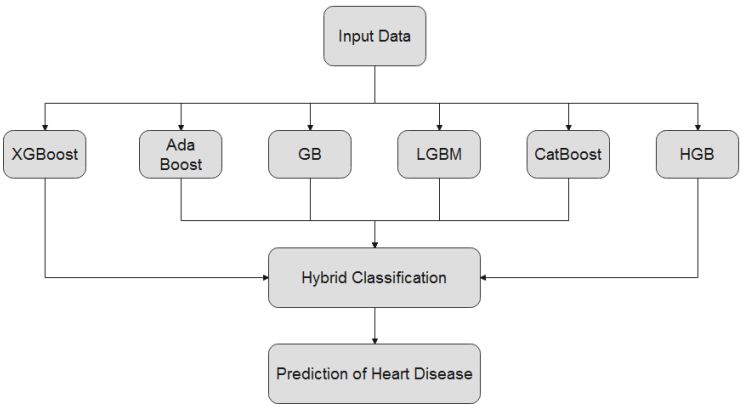
Table 4. LDA accuracy-test wall time

Accuracy on the test set	Wall time
0.8032786885245902	960 ms
0.8524590163934426	452 ms
0.8524590163934426	468 ms

Normally, the weak learner is utilised as the base classifier in classifier ensembles, however in this paper, we used strong ensemble learners as the base classifiers. Grid search is used to find the optimum learning hyperparameters for each base classifier by testing out all the available values. The AUC works here as a stopping metric of the search process. The classification of algorithms is shown in Fig. 12.

Hybrid classification technique that combines basic classification algorithms for model induction and data preprocessing. Misclassification instances are usually considered to be noise, to carry useful information for identifying the class values of some other instances. Meta classifier makes a final prediction among all the predictions by using predictions as features. Hybrid machine learning models essentially feed their output to one another (one-way) to create an efficient and accurate machine learning model.

Figure 12. Classification of boosting algorithms



4.6 Training Model

There are six different types of credit rating algorithms used in this paper. They are XGBoost, AdaBoost, Gradient Boosting, Light GBM, CatBoost, and Histogram-based Gradient Boosting. Using these algorithms six different predictive models were implemented in prediction probable phase-I. Using these algorithms, a hybrid ensemble technique is proposed. A hybrid ensemble technique is a combination of more than one base-level classifier. For classifying the heart disease, Bagging and Boosting techniques are applied. The classification process is used for labeling the dataset, which contains various numerical and nominal values, by constructing a model for operating various operations. The classification algorithms produce false leading due to noise data. Ensemble methods can resolve the false learning issues. In the process of hybrid ensemble machine learning, we identify one Meta classifier algorithm from each group of classification algorithms with the help of particle swarm optimization (PSO) and voting technique. There are several algorithms used for classification problems and pattern recognition.

5. RESULT AND DISCUSSION

In this paper, a new model "Intelligent heart Disease Prediction Empowered with novel boosting machine learning algorithms" is proposed for heart disease diagnosis with better accuracy. The heart disease dataset is used by different novel boosting algorithms like XGBoost, AdaBoost, Gradient Boosting, Light GBM, CatBoost, and Histogram-based Gradient Boosting. Furthermore, for splitting

the dataset into different folds for training and testing purpose the k-fold cross-validation method is used, By experiments, it's proved the proposed model accurately predicted heart disease with higher precision, and accuracy.

The accuracy is calculated through the following equation (5)

$$\text{Accuracy} = \frac{TR_{Positive} + TR_{Negetive}}{TR_{Positive} + TR_{Negetive} + FA_{Positive} + FA_{Negetive}} \times 100 \tag{5}$$

Precision calculated through the following equation (6)

$$\text{Precision} = \frac{TR_{Positive}}{TR_{Positive} + FA_{Positive}} \tag{6}$$

Recall calculated through the following equation (7)

$$\text{Recall} = \frac{TR_{Positive}}{TR_{Positive} + FA_{Negetive}} \tag{7}$$

Finally, the F-Measure can be calculated through the following equation(8)

$$\text{F1\_score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{8}$$

The Proposed system predicts the output value for positive $\left(+1\right)$ and negative $\left(-1\right)$ respectively. Where $+1$ represents the presence of heart disease and -1 represents the absence or no symptom of heart disease found in the patient body. The forecast performance of the proposed supervised ensemble-based boosting machine learning algorithms with Meta classifier statistical metrics is shown in Table 7. The training accuracy, testing accuracy, and accuracy difference of all the six ensemble-based boosting classifiers are shown in Table 5. It is clearly shown that the proposed architecture achieved an effective and highest accuracy results for XGBoost among the other six boosting algorithms with 96.33%, 94.18%, 2.15% for Training Accuracy, Testing Accuracy, and Accuracy difference respectively. Fige 13 and Fig 14 presents the graphical box plot comparison among boosting machine learning approaches of the base level classifiers for training accuracy.

5.1 Performances of Fusion-Based Meta-Classifiers

In this section, we will find the best combination of Meta Classifiers. The various types of Meta classifiers are used in our proposed technique are shown in Table 6. For developing the meta-classifier, we have used the stacking cross-validation technique to combine level-1 predictive output performance and finally develop a Meta classifier. The graphical Comparison box plots for the construction of a predictive Meta classifier without Cross-validation is shown in Fig 15(a), 15(b), 15(c),15(d),15(e),15(f).

The graphical Comparison box plots for Construction of predictive Meta classifier with Cross-validation is shown in Fig 16(a),16(b),16(c),16(d),16(e).

In our proposed technique 6 different Meta classifiers i.e XGBoost, AdaBoost, GB, LightGBM, CatBoost, HistoGradient Boosting are used, which are shown in Table 7. Out of these Meta classifiers,

Table 5. Forecast performances of the base level classifiers

	XGBoost	AdaBoost	Gradient Boost	LGBM	CatBoost	Histogram-based Boost
Training accuracy	96.33	95.7	94.3	95.27	94.93	95.42
Testing accuracy	94.18	93.99	92.6	92.73	92.99	92.09
Accuracy difference	2.15	1.71	1.7	2.54	1.94	3.33

Figure 13. The box plot comparison among boosting machine learning approaches of the base-level classifiers

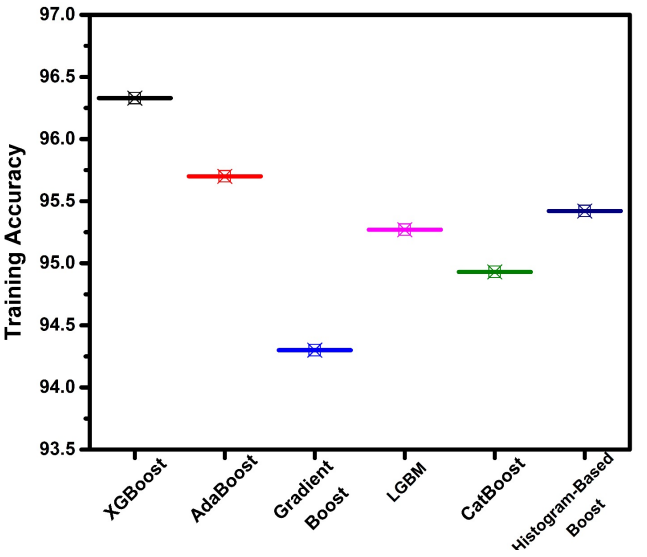


Figure 14. The box plot comparison among boosting machine learning approaches of the base-level classifiers

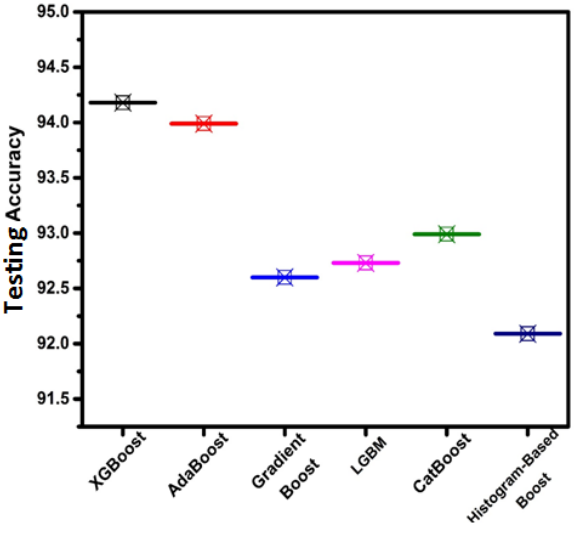


Table 6. Construction of predictive meta classifier without cross-validation

Models	Training Accuracy	Testing Accuracy	Accuracy	Difference
Meta-XGBoost	96.51	93.37	93.37	3.14
Meta-AdaBoost	96.33	94.38	93.28	1.95
Meta-G.B	96.48	94.48	93.18	2
Meta-LightGBM	95.5	94.28	93.28	1.22
Meta-CatBoost	96.38	94.18	93.28	2.2
Meta-H.G.boost	95.48	93.37	93.37	2.11

Figure 15. (a), 15(b), 15(c), 15(d), 15(e), 15(f) The graphical box plot comparison for construction of predictive meta classifier without cross-validation

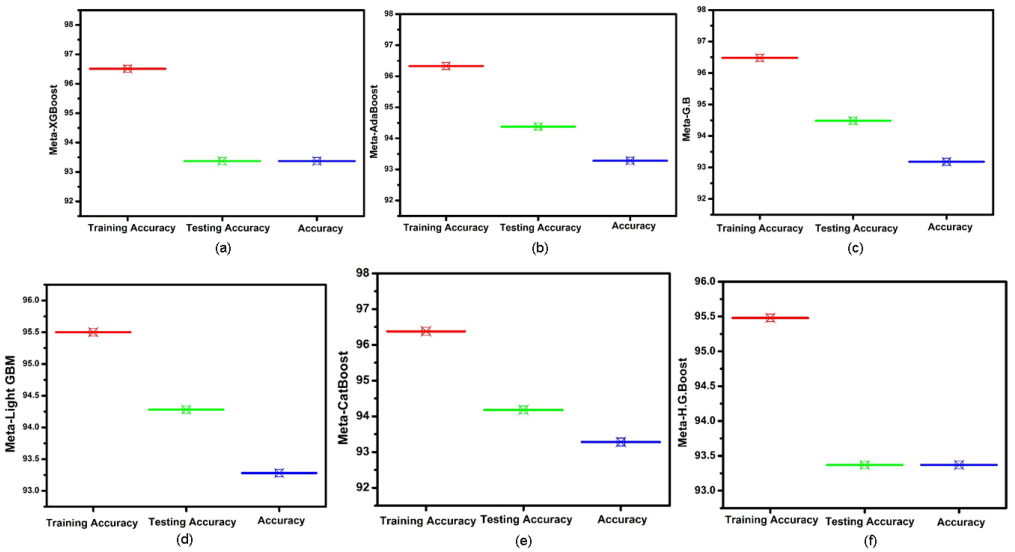


Table 7. Construction of predictive meta classifier with cross-validation

Models	Training Accuracy	Testing Accuracy	Accuracy	Difference
Meta-XGBoost	93.99	92.99	93	1
Meta-AdaBoost	94.23	92.71	92.72	1.52
Meta-GB	94.55	93.09	93.09	1.46
Meta-Light GBM	92.33	92.27	93.28	0.06
Meta-CatBoost	94.56	92.89	93	1.67
Meta-H.G.Boost	94.76	93.56	93.46	1.2

Figure 16. (a), 16(b), 16(c), 16(d), 16(e), 16(f) The graphical comparison for construction of predictive meta classifier with cross-validation

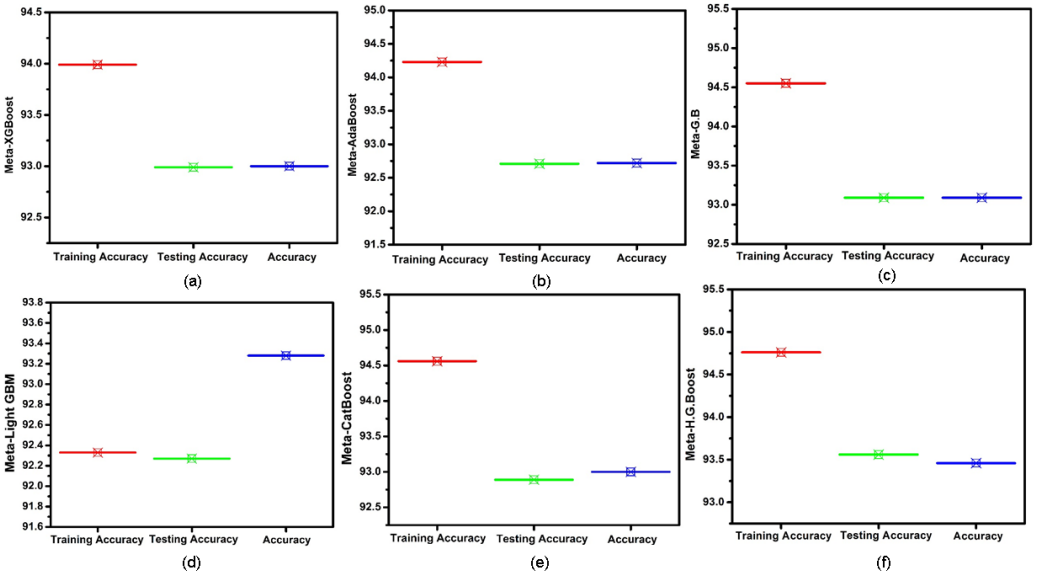


Table 8. Classification report meta-LightGBM

Precision	Recall	F1_Score	Accuracy
0	0.94	0.93	0.94
1	0.95	0.94	0.94

Table 9. Classification report meta-HGB

Precision	Recall	F1_Score	Accuracy
0	0.95	0.93	0.95
1	0.94	0.96	0.93

some of them have good Accuracy some of them are better but the LightGBM and Histogram-based Gradient Boosting give a good promising result as they trend to prove themselves as a generalized model with good predictive power. The model is generalized when the training accuracy and testing accuracy results are merely nearer to each other with a small difference or no difference.

The performance analysis results of different machine learning algorithms including XGBoost, AdaBoost, Gradient Boosting, Light GBM, CatBoost, and Histogram-based Gradient Boosting, and their comparison with our proposed heart disease prediction permits with fusion-based Meta classifier algorithm. The performance evaluation is evaluated using the parameters of classification accuracy that included computing of True Positive, False Positive using a comparison graph for them. Moreover, A ROC analysis is performed to identify the results and real-time facts. The Classification matrix of Meta LightGBM and Meta HGB are shown in Table 8 and Table 9.

Table 5 to Table 9 shows the comparison of metrics like Accuracy, Training Accuracy, Testing Accuracy, Difference, Recall, F1_Score, and the overall comparison of Meta classifier with other models. From the observations, it is known that the Meta-XGBoost Classifier without cross-validation model outperforms other approaches and gives better performance. The Meta HG Boost with cross-validation model outperforms other approaches and gives better performances. The accuracy is 96.51% higher than other Meta classifiers shows better performance in contrast to other approaches using without cross-validation and the accuracy is 94.76% higher than other Meta classifier shows better performance in contrast to other approaches using cross-validation. Thereby, the prediction of heart disease using Meta HG Boost helps the physicians to take an appropriate decision during the time of critical condition and acts as a better heart disease prediction.

The proposed model is compared with the recently published research model in terms of the existing performance and its accuracy. It has been proved that the proposed model produced higher accuracy compared to the previous research models. The comparative results from various existing models are shown in Table 10.

Table 10. Comparison of the proposed model with literature survey models

Sl. No	Author	Year	Model	Accuracy
1	Kahramanli, H., et al.	2008	Hybrid neural network for classification of data using FNN2 type and ANN	84.24%
2	Das, R., et al.	2009	A hybrid technique called neural network ensemble using SAS enterprise miner 5.2 programs to analyze and recognize heart disease by combining several neural networks with ensemble node	91.65%
3	El Bialy, et al.	2016	Ensemble methods for the support vector machine using Bayes Net, Multilayer perceptron, Nave Bayes, Decision tree classifiers, and Sequential Minimal Optimization for classification algorithms such as Bayes Net, Multilayer perceptron, Nave Bayes, Decision tree classifiers, and Sequential Minimal Optimization	90.00%
4	Mohan, S., et al.	2019	HRFLM (Hybrid Random Forest with Linear Model) is a Random Forest classification technique that uses Hybrid Random Forest with Linear Model (HRFLM).	88.70%
5	Latha, C. B. C., et al.,	2019	An ensemble classification model using Naïve Bayes, Random Forest, Bayes Net, C4.5, multilayer perceptron, and PART	83.17%
6	T.A. Asfaw., et al.	2019	K-Nearest Neighbors and Gaussian Nave Bayes classification algorithms are used to forecast heart disease.	77.05%
7	Yang, L. et al.	2020	CART, Nave Bayes, Bagged Trees, and Ada Boost are used in a random forest-based cardiovascular disease prediction model.	71.43%
8	Hashi, E. K., et al.	2020	Logistic regression, K-nearest neighbor, Support vector machine, Decision tree, and Random Forest are used in a heart disease prediction model.	90.16%
9	Almustafa, K. M. et al.	2020	K-Nearest Neighbor (K-NN), Naive Bayes, Decision Tree J48, JRip, SVM, Adaboost methods, and Stochastic Gradient Decent (SGD), Decision Table (DT) classifiers were used to classify the HD dataset.	93.85%
10	Budholiya, K. et al.	2020	Using XGBoost, Random Forest(RF), and Extra Tree(ET) classifiers, an efficient XgBoost classifier for heart disease prediction was developed.	91.80%
11	Khan M. A. et al.	2020	A cloud-based, intelligent system for predicting heart illness	93.33%
12	Nasarian, E., et al.	2020	Decision tree (DT), Gaussian Naive Bayes (GNB), Random Forest (RF), and XGBoost classifiers are used in a hybrid feature selection model.	81.23%
13	Tama, B. A., et al.	2020	Model for detecting coronary heart disease (CHD) combining random forest, gradient boosting machine, and extreme gradient boosting	83.90%
14	Karthikeyan, G., et al.	2021	A Hybrid Linear stacking model using XGBoost, HLS-XGBoost	96.00%
Proposed model				
15	A hybridization of fusion-based ensemble model with machine learning approaches including XGBoost, AdaBoost, Gradient Boosting, LightGBM, CatBoost, and Histogram-based Gradient Boosting meta classifier with cross-validation.			94.76%
16	A hybridization of fusion-based ensemble model with machine learning approaches including XGBoost, AdaBoost, Gradient Boosting, LightGBM, CatBoost, and Histogram-based Gradient Boosting meta classifier with out cross-validation.			96.51%

6. CONCLUSION AND FUTURE WORK

In this paper, a new hybrid model for heart disease prediction is proposed. The anticipated hybridization Meta classifier model is used for predicting the optimal feature subset and the Meta classifier algorithms for predicting the disease. The proposed model helps to address various problems like over-fitting and under-fitting issues while working with healthcare data from the available dataset. The Meta Histogram-based Gradient Boosting model is compared with other Meta classifier algorithms like XGBoost, AdaBoost, GB, LightGBM, and CatBoost. An extensive analysis performed by evaluating the results of the proposed and existing approaches i.e Accuracy, Training Accuracy, Testing Accuracy, Difference, Recall, F1_Score, and the overall comparison of Meta classifier. It is proved the outperformance of the proposed Meta Histogram-based Gradient boosting model gives higher accuracy than any other ML approach. The comparison of various feature selection approaches with Histogram-based Gradient boosting and the classification model enhances the prediction performance. From the experimental analysis, the Histogram-based Gradient boosting model enhances the prediction rate with an accuracy of 96.51% and assists the physicians to formulate an effectual decision during the complication period.

We plan to introduce an efficient Remote heart disease prediction system in the future, which will monitor and predict heart disease using patient data obtained from remote devices, and new algorithms can be included to the hybrid ensemble classifier to increase its performance.

COMPETING INTERESTS

The authors have declared that there is no conflict of interest regarding the publication of this paper.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

Al Rahhal, M. M., Bazi, Y., AlHichri, H., Alajlan, N., Melgani, F., & Yager, R. R. (2016). Deep learning approach for active classification of electrocardiogram signals. *Information Sciences*, *345*, 340–354. doi:10.1016/j.ins.2016.01.082

Al-Shayea, Q. K. (2011). Artificial neural networks in medical diagnosis. *International Journal of Computer Science Issues*, *8*(2), 150–154.

Almustafa, K. M. (2020). Prediction of heart disease and classifiers' sensitivity analysis. *BMC Bioinformatics*, *21*(1), 1–18. doi:10.1186/s12859-020-03626-y PMID:32615980

Ansarullah, S. I., & Kumar, P. (2019). A systematic literature review on cardiovascular disorder identification using knowledge mining and machine learning method. *Int. J. Recent Technol. Eng.*, *7*(6S), 1009–1015.

Ardakany, A. R., Naderi, E., & Osareh, A. (2010, December). Parallel weak learners, a novel ensemble method. In *2010 IEEE International Conference on Computational Intelligence and Computing Research* (pp. 1-4). IEEE. doi:10.1109/ICCIC.2010.5705773

Asfaw, T. (2019). Performance comparison of k-nearest neighbors and Gaussian naïve bayes algorithms for heart disease prediction. *Int. J. Eng. Sci. Invent.*, *8*(8), 45–48.

Budholiya, K., Shrivastava, S. K., & Sharma, V. (2020). An optimized XGBoost based diagnostic system for effective prediction of heart disease. *Journal of King Saud University-Computer and Information Sciences*.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). doi:10.1145/2939672.2939785

Coats, A. J. S. (2019). Ageing, demographics, and heart failure. *European Heart Journal Supplements*, *21*(Supplement_L), L4–L7. doi:10.1093/eurheartj/suz235 PMID:31885504

Das, R., Turkoglu, I., & Sengur, A. (2009). Effective diagnosis of heart disease through neural networks ensembles. *Expert Systems with Applications*, *36*(4), 7675–7680. doi:10.1016/j.eswa.2008.09.013

Drucker, H., Schapire, R., & Simard, P. (1993). Boosting performance in neural networks. In Advances in Pattern Recognition Systems using Neural Network Technologies (pp. 61-75). doi:10.1142/S0218001493000352

Durairaj, M., & Ramasamy, N. (2016). A comparison of the perceptive approaches for preprocessing the data set for predicting fertility success rate. *Int. J. Control Theory Appl*, *9*(27), 255–260.

El Bialy, R., Salama, M. A., & Karam, O. (2016, May). An ensemble model for Heart disease data sets: a generalized model. In *Proceedings of the 10th international conference on informatics and systems* (pp. 191-196). doi:10.1145/2908446.2908482

Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Information and Computation*, *121*(2), 256–285. doi:10.1006/inco.1995.1136

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society for Artificial Intelligence, 14*(771-780), 1612.

Gonsalves, A. H., Thabtah, F., Mohammad, R. M. A., & Singh, G. (2019, July). Prediction of coronary heart disease using machine learning: An experimental analysis. In *Proceedings of the 2019 3rd International Conference on Deep Learning Technologies* (pp. 51-56). doi:10.1145/3342999.3343015

Guidi, G., Pettenati, M. C., Melillo, P., & Iadanza, E. (2014). A machine learning system to improve heart failure patient assistance. *IEEE Journal of Biomedical and Health Informatics*, *18*(6), 1750–1756. doi:10.1109/JBHI.2014.2337752 PMID:25029521

Hashi, E. K., & Zaman, M. S. U. (2020). Developing a hyperparameter tuning based machine learning approach of heart disease prediction. *Journal of Applied Science & Process Engineering*, *7*(2), 631–647. doi:10.33736/jaspe.2639.2020

Heidenreich, P. A., Trogdon, J. G., Khavjou, O. A., Butler, J., Dracup, K., Ezekowitz, M. D., Finkelstein, E. A., Hong, Y., Johnston, S. C., Khera, A., Lloyd-Jones, D. M., Nelson, S. A., Nichol, G., Orenstein, D., Wilson, P. W. F., & Woo, Y. J. (2011). Forecasting the future of cardiovascular disease in the United States: A policy statement from the American Heart Association. *Circulation*, *123*(8), 933–944. doi:10.1161/CIR.0b013e31820a55f5 PMID:21262990

Kahramanli, H., & Allahverdi, N. (2008). Design of a hybrid system for the diabetes and heart diseases. *Expert Systems with Applications*, *35*(1-2), 82–89. doi:10.1016/j.eswa.2007.06.004

Karthikeyan, G., Komarasamy, G., & Daniel Madan Raja, S. (2021). An Efficient Method for Heart Disease Prediction Using Hybrid Classifier Model in Machine Learning. *Annals of the Romanian Society for Cell Biology*, 5708–5717.

Kearns, M. (1988). *Learning Boolean formulae or finite automata is as hard as factoring*. Technical Report TR-14-88 Harvard University Aikem Computation Laboratory.

Kearns, M., & Valiant, L. (1994). Cryptographic limitations on learning Boolean formulae and finite automata. *Journal of the Association for Computing Machinery*, *41*(1), 67–95. doi:10.1145/174644.174647

Khan, M. A., Abbas, S., Atta, A., Ditta, A., Alquhayz, H., Khan, M. F., & Naqvi, R. A. (2020). *Intelligent cloud based heart disease prediction system empowered with supervised machine learning*. Academic Press.

Khare, P., & Burse, K. (2016). Feature selection using genetic algorithm and classification using weka for ovarian cancer. *International Journal of Computer Science and Information Technologies*, *7*(1), 194–196.

Lahner, E., Intraligi, M., Buscema, M., Centanni, M., Vannella, L., Grossi, E., & Annibale, B. (2008). Artificial neural networks in the recognition of the presence of thyroid disease in patients with atrophic body gastritis. *World Journal of Gastroenterology: WJG*, *14*(4), 563. doi:10.3748/wjg.14.563 PMID:18203288

Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, *16*, 100203. doi:10.1016/j.imu.2019.100203

Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A., & Saboor, A. (2020). Heart disease identification method using machine learning classification in e-healthcare. *IEEE Access: Practical Innovations, Open Solutions*, *8*, 107562–107582. doi:10.1109/ACCESS.2020.3001149

Melillo, P., De Luca, N., Bracale, M., & Pecchia, L. (2013). Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability. *IEEE Journal of Biomedical and Health Informatics*, *17*(3), 727–733. doi:10.1109/JBHI.2013.2244902 PMID:24592473

Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access: Practical Innovations, Open Solutions*, *7*, 81542–81554. doi:10.1109/ACCESS.2019.2923707

Mohapatra, S., Patra, D., & Satpathy, S. (2014). An ensemble classifier system for early diagnosis of acute lymphoblastic leukemia in blood microscopic images. *Neural Computing & Applications*, *24*(7), 1887–1904. doi:10.1007/s00521-013-1438-3

Nahar, J., Imam, T., Tickle, K. S., & Chen, Y. P. P. (2013). Association rule mining to detect factors which contribute to heart disease in males and females. *Expert Systems with Applications*, *40*(4), 1086–1093. doi:10.1016/j.eswa.2012.08.028

Nasarian, E., Abdar, M., Fahami, M. A., Alizadehsani, R., Hussain, S., Basiri, M. E., Zomorodi-Moghadam, M., Zhou, X., Pławiak, P., Acharya, U. R., Tan, R.-S., & Sarrafzadegan, N. (2020). Association between work-related features and coronary artery disease: A heterogeneous hybrid feature selection integrated with balancing approach. *Pattern Recognition Letters*, *133*, 33–40. doi:10.1016/j.patrec.2020.02.010

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

Ray, P., Kharke, R. B., & Chauhan, S. S. (2021). Cardiovascular Disease Classification Using Different Algorithms. In *Inventive Communication and Computational Technologies* (pp. 189–201). Springer.

Sarwar, A., Sharma, V., & Gupta, R. (2015). Hybrid ensemble learning technique for screening of cervical cancer using Papanicolaou smear image analysis. *Personalized Medicine Universe*, *4*, 54–62. doi:10.1016/j.pmu.2014.10.001

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, *5*(2), 197–227. doi:10.1007/BF00116037

Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computers and Applications*, *17*(8), 43–48. doi:10.5120/2237-2860

Spoletini, I., & Seferovic, P. (2017, June). The Management of Co-Morbidities in Patients with Heart Failure–Angina and Coronary Disease. In *International Cardiovascular Forum Journal* (Vol. 10). doi:10.17987/icfj.v10i0.451

Tama, B. A., Im, S., & Lee, S. (2020). Improving an intelligent detection system for coronary heart disease using a two-tier classifier ensemble. *BioMed Research International*, *2020*, 2020. doi:10.1155/2020/9816142 PMID:32420387

Tsanas, A., Little, M. A., McSharry, P. E., & Ramig, L. O. (2011). Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity. *Journal of the Royal Society, Interface*, *8*(59), 842–855. doi:10.1098/rsif.2010.0456 PMID:21084338

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, *27*(11), 1134–1142. doi:10.1145/1968.1972

Woźniak, M., Grana, M., & Corchado, E. (2014). A survey of multiple classifier systems as hybrid systems. *Information Fusion*, *16*, 3–17. doi:10.1016/j.inffus.2013.04.006

Yang, L., Wu, H., Jin, X., Zheng, P., Hu, S., Xu, X., Yu, W., & Yan, J. (2020). Study of cardiovascular disease prediction model based on random forest in eastern China. *Scientific Reports*, *10*(1), 1–8. doi:10.1038/s41598-020-62133-5 PMID:32251324

Zhang, R., Ma, S., Shanahan, L., Munroe, J., Horn, S., & Speedie, S. (2017, November). Automatic methods to extract New York heart association classification from clinical notes. In *2017 ieee international conference on bioinformatics and biomedicine (bibm)* (pp. 1296-1299). IEEE.

Zhao, L., Ni, L., Hu, S., Chen, Y., Zhou, P., Xiao, F., & Wu, L. (2018, April). Inprivate digging: Enabling tree-based distributed data mining with differential privacy. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications* (pp. 2087-2095). IEEE.

Zheng, H., Yuan, J., & Chen, L. (2017). Short-term load forecasting using EMD-LSTM neural networks with a Xgboost algorithm for feature importance evaluation. *Energies*, *10*(8), 1168. doi:10.3390/en10081168

Sibo Prasad Patro received his MCA in 2012 from Sambalpur University, Odisha and his M.Tech (Computer Science and Engineering) in 2014 from BPUT, Raurkela, Odisha. He is currently pursuing his Ph.D in Computer Science and Engineering at GIET University, Gunupur, Odisha under the supervision of Dr. Neelamadhab Padhy and Dr. Rahul Deo Sah. He has published 1 SCI indexing as well as Scopus indexing file in Springer. He has also published few conference papers. He has more than 15 years of teaching experience. His research interest includes machine learning, data mining, IoT and their application to engineering. Currently he is working as Assistant Professor in the department of Computer Science and Engineering, GIET University, Gunupur. He has received a best paper presentation award in ICCSEA-2020 an International Conference organized by GIET University, Gunupur.

Neelamadhab Padhy received his MCA in 2003 from the BPUT, Rourkela, and his M.Tech(CS) in 2009 form the Berhampur University, Odisha. He received his Ph.D. in Computer Science and Engineering from SSSUTM, Sehore, under the supervision of Prof. R.P.Singh and Prof. Suresh Chandra Satapathy. He has published many articles in SCI indexing as well as Scopus Indexing like Springer, Elsevier and Inderscience etc. He has also published many conference paper and book chapters. He has ore then 50 publications to his credit in various reputed international journals and conference proceedings. He has more than 17 years of teaching and research experience. His research interests include machine learning, data mining, swarm intelligence studies and their application to engineering. Recently he has published 5 patents and copyright. Currently he is working as an Associate Professor in GIET University, Gunupur. His Google scholar citation reached to 592, H-Index is 12 and i-10 index is 13. He served reviewer and editorial board member of the several journal and conference.

Rahul Deo Sah got his Ph.D. in Computer Sc. and Application. Currently Dr Sah is working as Assistant Professor in Dept. Computer Application and Information Technology, Dr Shyama Prasad Mukherjee University Ranchi Jharkhand. Sah is a Life member of Computer Society of India, and Indian Science Congress, India. He is the Member of Research Foundation of India.Dr. Sah is also the RFI chapter Head of Jharkhand State.Dr. Sah a Life member of InSc, India. He has authored around twenty Five articles in different International/National journals and conferences. He has guided UG, PG and Research Scholars.Dr. Sah has Editor of one Book. He has Authored 02 books chapter in Springer. Dr Sah got International Award for Think Tank in Technical Education on 3rd January 2020. Dr. Sah has been selected for The World's Prestigious "Global Educational Award 2020" on Teacher's Day Special Award - 2020 - BEST ACADEMICIAN for their outstanding excellence and remarkable achievements in the field of Teaching, Research and Publications by ESN Publications, Chennai. He is a coordinator of STP IIT BOMBAY,Dr. SPM University Ranchi and that is also resource center of the university. He has given talk in different Workshops / FDPs / Seminars. Dr Sah has organized many Workshops at university level.Dr. Sah has been chaired Special Sessions in different IEEEs and Springer International/National Conferences.