International Journal on Engineering Technology (InJET)

www.kec.edu.np/journal
Volume 1, issue no. 1, Nov 2023, Page 83-101
Received Date: Sept 3, 2023
Accepted Date: Nov 2, 2023
ISSN: 3021-940X (print)

Heart Disease Prediction Using Outlier Removal based Max Voting Ensemble Method

Pralhad Chapagain*

Kantipur Engineering College, Dhapakhel, Lalitpur Nepal, pralhadchapagain@kec.edu.np

Abstract:

Heart disease has emerged as a serious health concern for many individuals due to its high death rate around the world. The routine clinical data analysis has a significant difficulty in the early diagnosis of cardiac disease. The identification of cardiac disease may benefit from the use of machine learning. To improve machine learning models, several studies have previously been conducted. The suggested study uses the maximum voting ensemble technique of classification to effectively identify heart disease. The suggested classifier is a more reliable and accurate approach. To identify and eliminate outliers, conduct Inter quartile range outlier removal and min-max normalization during preprocessing. Accuracy, Precision, Recall, and F1 Score are calculated and evaluated against various models. For the heart disease dataset collected from the Kaggle, the suggested max voting ensemble classifier has an accuracy of 99.22%.

Keywords: Heart Disease, Max Voting, Ensemble, Outlier Removal, XGBOOST, Decision Tree, KNN, SVM, Gradient Boost

1. Introduction:

There are several problems that affect the heart, including coronary artery disease, arrhythmia, and heart failure, collectively referred to as heart disease. Today's generation is exceedingly active in their daily routine, which can lead to feelings of anxiousness and stress. Lifestyle choices can prevent many of these conditions, and medications can help manage them if they arise. Each individual has unique heart rate and blood pressure levels. Pulse rates typically vary between 60 to 100 beats per minute, while blood pressure usually falls within the range of 120/80 to 140/90. Globally, Heart Disease (HD) stands as the primary contributor to mortality. (Cenitta, et al., 2022)

Heart failure is a state wherein the heart cannot effectively pump sufficient blood to satisfy the body's requirements. It is primarily linked to the heart's structure, such as a past heart attack, and secondarily to its function, like elevated blood pressure. Indications of heart failure encompass breathlessness, weariness, and edema in the legs and ankles. Approaches for addressing heart failure comprise medications, alterations in lifestyle, and occasionally, surgical procedures. Studies have demonstrated that identifying and addressing heart failure promptly can enhance life quality and extend survival rates. (Qadri, et al., 2023)

It is difficult to predict cardiac disease since it calls for both extensive expertise and cutting-edge information. (Khan, 2020) The severity of the heart disease problem is categorized using a variety of techniques, including the K-Nearest Neighbor Algorithm (KNN), Naive Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), and Genetic Algorithm (GA). Given the complexity of the condition, treating heart disease requires caution. Failure to comply might cause death or damage to the heart. To identify many types of metabolic problems, medical

*Corresponding Author

research and machine learning (ML) are being used. Machine learning with categorization plays a vital role in the classification of heart disease and data analysis. (Mohan, et al., 2019)

The most common conditions nowadays that shorten people's lifespans are heart illnesses. The World Health Organization estimates that 17.9 million people die each year. In recent years, the global epidemic of cardiovascular infection has been growing swiftly on a global scale. So, the main risk to save their lives is anticipating a heart infection. (World Health Organization, 2023)The research focus on the requirement to predict if individuals have cardiovascular sickness by providing a few client characteristics. It matters for clinical fields. If an expectation is sufficiently accurate, we can prevent incorrect analysis and even save lives. We can address the unnecessary problems if we have realistic expectations. Furthermore, by eliminating the need for complicated detection processes in medical clinics, machine learning in clinical forecasting would save human wealth. (Verma & Gupta, 2021)

The major challenge lies in the substantial volume of patient data within the medical domain, which often remains underutilized by analysts and professionals. The clinical data pertaining to patients contains latent patterns that hold great significance for information analysis in the context of diagnosing coronary disease. Data mining assumes a pivotal role in deciphering diverse information within this specialized field, enabling effective analysis and interpretation. The dataset on coronary disease, accessible from the UCI repository, can be transformed into meaningful insights through the application of data mining techniques. This process involves uncovering concealed patterns and insights by analyzing a substantial volume of data stored within the dataset, using methods like machine learning. Thus, this research selectively gathers pertinent information related to the domain of study, train the data using various machine learning algorithms, and predict the likelihood of heart disease occurrence in its early stages. (Hazra, et al., 2017)

In this research, the UCI (University of California Irvine) heart disease dataset was sourced from Kaggle. The initial step involved data preprocessing, wherein outliers were identified and removed using the interquartile range method. Subsequently, normalization was carried out to standardize the data. Once the preprocessing was completed, an ensemble model was trained. Various parameters, such as accuracy, Recall, precision and F1 are meticulously analyzed to evaluate the performance of the trained model. This comprehensive approach ensures the reliability and effectiveness of the ensemble model in predicting heart disease outcomes.

2. Related works

The most important area for machine learning applications is heart disease. Therefore, various studies have been conducted in the past to precisely predict heart disease. A select few of these are discussed in this literature study.

(Raza, 2019) did research on how to use ensemble learning and the majority voting rule to increase the forecast accuracy of heart disease. The majority vote rule is used to merge machines learning techniques logistic regression, multilayer perceptron, and naive bays. The suggested model's accuracy was 88.88% using data from the UCI repository.

(Atallah & Al-Mousa, 2019) proposed a Heart Disease Prediction using Machine learning Majority Voting Ensemble Method. Aim of the research is to provide more confidence and accuracy to the Doctor’s diagnosis since model is trained using real life data of healthy and ill patients. The proposed model classifies the patient based on majority vote of several machine learning models: Stochastic Gradient Descent (SGD) Classifier, K-Nearest Neighbor Classifier, Random Forest Classifier and Logistic Regression Classifier in order to provide more accurate solution than having only single model. The presented model provides the accuracy of about 90% on UCI repository Cleveland Database.

(Rafsun, et al., 2022) introduced an ensemble architecture to detect and analyze the heart disease in their research paper Heart Disease Prediction and Analysis using Ensemble Architecture. Ensemble architecture could help weak algorithms increase their performance. In the study, they offered a unique ensemble architecture that uses a hard voting mechanism to improve the performance. XGBoost, Logistic Regression, Random Forest and K-Nearest Neighbor algorithms are employed in the ensemble architecture. This model scores were obtained with a 94% accuracy rate on dataset available in UCI repository.

(Madhu & Ramesh, 2021) done their research in Heart Attack Analysis and Prediction using SVM where author used UCI health care heart disease data set to evaluate the model. The proposed model yields the 83% of accuracy.

(Sumwiza, et al., 2023) used the outlier removal in their research paper Enhanced cardiovascular disease prediction model using random forest algorithm. Correlation coefficient and data mining feature selection techniques are applied to remove the outlier. After that random forest is used to train the model and dataset from the Kaggle repository with 1025 data points and 14 attributes are used to evaluate the system. After employed the feature selection algorithm the accuracy of the random forest is almost 99% is obtained.

(Singh & Kumar, 2020) Conducted the analysis of heart disease prediction employing diverse machine learning methodologies. Within this study, classification and regression models are harnessed for predictive purposes, encompassing the Decision Tree, K-Nearest Neighbors (KNN) algorithm, Support Vector Machine (SVM), and linear regression approaches. The empirical findings derived from experiments showcased the KNN algorithm as yielding the most robust accuracy. Nonetheless, it is worth noting that this model holds potential for real-time implementation within practical environments or applications.

(Almazroi, et al., 2023), introduced a clinical decision support system for predicting heart disease using deep learning. In this research, a dense neural network was applied to various datasets, and the model's performance was assessed. The proposed model demonstrated superior accuracy when compared to other machine learning algorithms, such as nearest neighbor, linear SVM, Gaussian process, decision tree, Naïve Bayes, and several others.

In reference to (Cenitta, et al., 2022), conducted a study on the prediction of ischemic heart disease through the application of the optimized squirrel search feature selection algorithm. The primary emphasis of the researchers lay in the realm of feature selection employing optimized algorithms. For prediction purposes, a classification algorithm, specifically the random forest, was utilized. The dataset under scrutiny was sourced from the UCI repository and Kaggle platforms. Within the dataset, a total of 76 features were present, out of which 14 exhibited considerable efficacies in discerning heart disease. The proposed model exhibited an accuracy exceeding 98%, a notable achievement. This level of accuracy was then juxtaposed against the performance of other machine learning algorithms, including naïve Bayes, k-nearest neighbors (KNN), and numerous others, for comparative analysis.

(Beunza, et al., 2019) in their research work employed the array of distinct machine learning classifiers, encompassing decision tree, random forest, support vector machines (SVM), neural networks, and logistic regression (LR) on dataset obtained from the UCI repository. The findings revealed that SVM emerged as the optimal classifier model, displaying a noteworthy area under the curve (AUC) value of 0.75.

Adequate research has been conducted in the realm of heart disease prediction using machine learning techniques. While many studies have primarily focused on algorithms for classification and feature selection, limited attention has been directed towards thoroughly cleaning the dataset and presenting a well-preprocessed dataset to the model, thereby bolstering its performance. In light of this, the proposed approach employs interquartile range-based outlier detection and removal techniques for dataset preprocessing. For classification purposes, a majority vote-based ensemble method is adopted. This involves integrating various machine learning classifiers, chosen based on their optimal accuracy as identified through an extensive literature review. This comprehensive methodology aims to enhance the overall performance of the proposed approach.

3. Methodology

3.1 Proposed Model

The architectural representation of the proposed method outlier removal-based ensemble method for heart disease prediction is illustrate in following diagram.

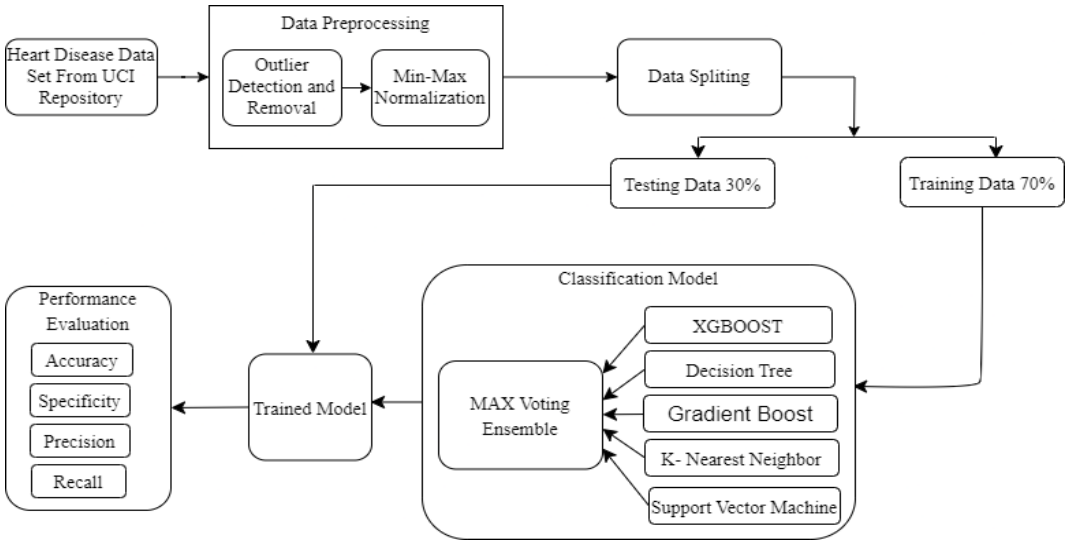

Figure 1:Architectural Representation of Proposed Model

The proposed architecture includes different steps. Firstly, heart disease dataset is collected from the Kaggle. The dataset is free from the null value so outlier detection and removal and normalization are done in preprocessing phase. After that the dataset is divided into training and testing dataset. The training data is used to train the proposed classification model and testing data is used to test and evaluate the performance of the system.

3.2 Data Preprocessing

Data preprocessing is a data mining technique that involves converting unstructured data into a usable form. In this research, the dataset from Kaggle is used and it has no null value. So, Outlier detection and removal and normalization is done in data preprocessing phase.

Outlier detection pertains to the task of identifying patterns within data that fall outside the scope of normal behavioral ranges. These aberrant patterns, known as outliers, deviate from the anticipated statistical distribution. The process of detecting and subsequently mitigating such outliers serves to enhance the precision of the system. Within each dataset, the presence of outliers is inevitable, arising from diverse causative factors. Several common rationales for their occurrence include:

1. Malicious activity
2. Instrumentation error
3. Change in the environment
4. Human error

Interquartile range (IQR) is a technique that helps to find outliers in the data which are continually distributed. It is the difference between the first quartile and third quartile is defined by equation 1.

$$\mathrm{IQR} = \mathrm{Q3} - \mathrm{Q1} \quad \text{(Equation 1)}$$

Where, IQR = Inter Quartile Range, Q_3 = Third quartile and Q_1=First Quartile

To detect the outlier lower and upper boundary are calculated by using equation 2 and 3.

$$\text{Lower Boundary} = \mathrm{Q1} - 1.5 * \mathrm{IQR} \quad \text{(Equation 2)}$$

$$\text{Upper Boundary} = \mathrm{Q3} - 1.5 * \mathrm{IQR} \quad \text{(Equation 3)}$$

The data resides outside of the lower and upper boundary are outliers and are removed to clean the dataset. (Vinutha, et al., 2018)

The objective of normalization is to systematically alter the values within numeric columns of a dataset in order to establish a uniform scale, all the while safeguarding the integrity of distinctions inherent in the value ranges and preventing information loss. Min-Max normalization is applied in this research and given by equation 4. (Chapagain, et al., 2022)

$$\text{Min} - \text{Max} = \frac{Value - Min\ Value}{Max\ Value - Min\ Value} \quad \text{(Equation 4)}$$

Where, Min-Max = Min-Max Normalization, Max. value = Maximum value and Min. value = Minimum Value

(Chapagain, et al., 2022)

3.4 Classification Model

The supervised machine learning paradigm of classification involves the model attempting to determine the correct label for an input dataset. The model is carefully trained on the training data within the classification domain, then subjected to assessment against test data, before being applied for predicting tasks requiring unique, unseen data points.

This study's use of ensemble learning techniques was a key component. A subfield of machine learning called ensemble learning uses the potential of integrating many models to boost prediction accuracy and robustness. In this research, for classification different classifiers, including K-Nearest Neighbor, Decision tree classifier, XGBoost, Support Vector Machine and Gradient Boost and combined using majority vote rule. The different algorithms for classification are chosen by doing literature review. Among the various classification approaches, the best one is chosen to develop the model. In different dataset and data preprocessing algorithms, the above-mentioned algorithms perform better according to the literature mentioned above.

In ensemble learning, many independent models are trained on the same dataset, and their individual predictions are then combined to generate a final prediction. An outline of this ensemble learning process, which combines the predictions from many classifiers to produce a single prediction, is shown in Figure 3. The use of performance indicators, such as but not limited to accuracy, precision, F1 and Sensitivity, makes it easier to evaluate ensemble models. (Asif, et al., 2023)

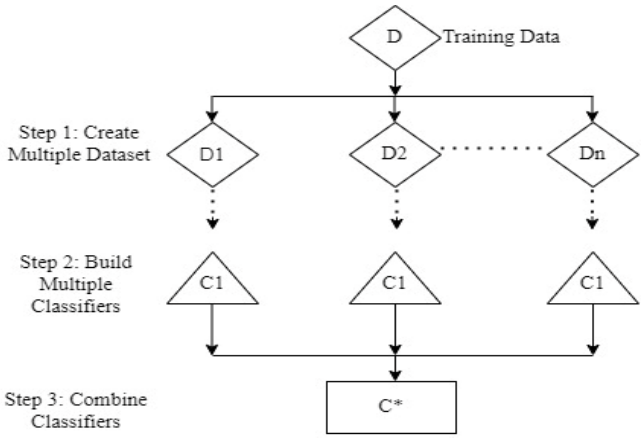

Figure 2: The ensemble learning procedure.

[Source: (Asif, et al., 2023)*]*

The individual algorithms used for combining using majority vote rule to form the proposed classifier model are discussed in detail.

3.4.1 Decision tree Classifier:

The decision tree constitutes a machine learning algorithm, functioning as both a regression and classification technique. Manifesting as a tree-like structure, it operates as a classifier. Illustrated in Figure 4, it constructs a

hierarchy of decision nodes, wherein internal nodes delineate data attributes, edges encapsulate decision rules, and each terminal node signifies a class label—specifically, the target class label should that path be traversed. Within a decision tree, the class prediction mechanism follows this progression: Commencing from the root node in adherence to decision rules, subsequent nodes are selected iteratively until a terminal node is reached. The assigned class label at the terminal node corresponds to the anticipated class label. (Saraswathi, et al., 2022)

Information gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute. It calculates how much information a feature provides about a class. According to the value of information gain the decision tree is built by splitting the node. A decision tree algorithm always tries to maximize the value of information gain, and a node/attribute having the highest information gain is split first. Information gain for attribute selection measures is given by the equation 5 and 6.

1. $Information\ Gain = Entropy(S) - [(Weighted\ Average) * Entropy(Each\ feature)]$ (Equation 5)
2. $Entropy(S) = -P(yes)\log_2(P(yes)) - P(no)\log_2(P(no))$ (Equation 6)
   $where,$
   $S = Total\ number\ of\ Data\ Sample$
   $P(Yes) = Probability\ of\ yes$
   $P(No) = Probability\ of\ No$

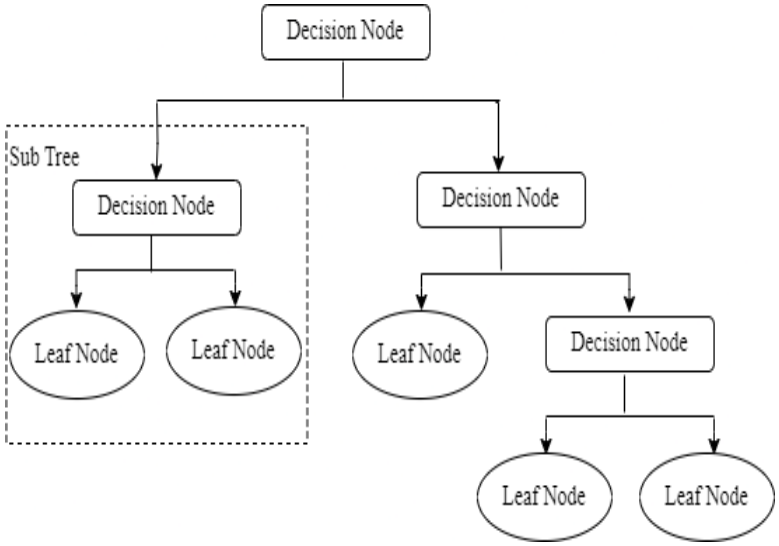

Figure 3: Decision Tree Model

3.4.2 K-Nearest Neighbor Classifier:

KNN, or k-nearest neighbors, stands as one of the most straightforward and uncomplicated data mining techniques. Referred to as Memory-Based Classification, KNN mandates the presence of training examples within memory during runtime.

The new unlabeled data is categorized using the closest neighbor approach by figuring out which category its neighbors fall within. This idea is incorporated into the KNN algorithm's computation. With the KNN method, a certain value of K is fixed, assisting us in categorizing the unknown tuple. (Taunk, et al., 2019)

The working of the algorithm is described by the following steps:

Step1: Store the training Set

Step2: for each new unlabeled data, calculate Euclidian distance with all training data points using the formula given in equation 7.

$$\sqrt{\sum_{i=1}^{n}(x_i - a)^2 + (y_i - b)^2} \quad \text{(Equation 7)}$$

$where,$
$n = total\ number\ of\ data\ set$
$x, y = training\ data\ set\ co-ordinate$
$a, b = coordinate\ of\ unlabeled\ data$
Step3: find the K-nearest neighbors

Step4: assign class containing the maximum number of nearest neighbors.

3.4.3 Support Vector Machine Classifier:

The Support Vector Machine (SVM) stands as a supervised machine learning algorithm with applications in both classification and regression tasks. While SVM is versatile enough to handle regression problems, its optimal utility lies in the domain of classification. The central objective of the SVM algorithm involves the identification of a hyperplane within an N-dimensional space that distinctly segregates the data points. The dimensionality of this hyperplane is contingent upon the quantity of input features. In scenarios where the input features number two, the hyperplane is represented as a line. For instances with three input features, the hyperplane assumes the form of a 2-dimensional plane. However, the visualization of hyperplanes becomes progressively intricate as the number of features surpasses three. (Madhu & Ramesh, 2021)

3.4.4 Gradient Boost Classifier:

The gradient boosting method employs a sequential approach, incrementally refining the algorithm based on a loss function. By detecting and correcting errors, it aims to enhance accuracy. Typically, boosting involves assessing models that minimize the loss function derived from trained data. These assessments lead to error quantification and analysis, essential for optimal result prediction. The loss function computes the extent of identified disparity, subsequently compared against the intended objective. The progressive stepwise approach, a prevalent technique, is employed to update models with varying attributes. Accuracy optimization is achieved by minimizing the loss function and introducing base learners across all stages. (Raja, et al., 2019)

Steps for gradient Boosting method

- Calculate the mean value of the target variable of the dataset.
- Determine the residuals for each data by applying equation 8.
    - $Residual = Actual\ Value - Predicted\ Value$ (Equation 8)
- Construct a decision tree aimed at forecasting the residuals.
- Forecast the target label using all ensemble trees with the assistance of equation 9.
    - $Predicted = Average\ of\ Target\ variable + Learning\ Rate * Residual\ Predicted\ by\ Decision\ Tree$ (Equation 9)

$where\ Learning\ Rate = Determines\ the\ contribution\ of\ each\ tree\ on\ the\ final\ outcome\ and\ controls\ how\ quickly\ the\ algorithm\ proceeds\ down\ the\ gradient\ descent\ , ranges\ from\ 0\ to\ 1$

- Compute the fresh residual using equation 8.
- Iterate through steps 3 to 5 until the iteration count matches the specified hyperparameter value (number of estimators).
- Once trained, use all of the trees in the ensemble to make a final prediction as to value of the target variable. The final prediction will be equal to the mean we computed in Step 1 plus all the residuals predicted by the trees that make up the forest multiplied by the learning rate.

3.4.5 XGBOOST:

XGBoost, which stands for Extreme Gradient Boosted trees, functions similarly to GBT in terms of its training process, but it incorporates a novel approach to tree construction. Unlike other ensemble algorithms, where trees are built using conventional methods such as Gini Impurity or Entropy, XGBoost introduces a fresh criterion termed "similarity score" for the selection and division of nodes. (Khan, 2021)The ensuing process of developing a Decision Tree through the utilization of similarity score encompasses the following steps:

- Make a tree with just one leaf.
- Calculate the average of the target variable as a forecast for the first tree, and then compute the residuals using the appropriate loss function. The residuals for succeeding trees originate from earlier tree predictions.
- Calculate the similarity score using the following equation 10:
    - $Similarity\ Score = \frac{Gradient^2}{Hessian + \lambda}$ (Equation 10)
    - Where, Hessian is equal to the number of residuals; Gradient^2 = squared sum of residuals; λ is a regularization hyperparameter.
- Choose the relevant node using the similarity score. The homogeneity is greater the higher the similarity score. Information gain is calculated using similarity score.
- Information gain indicates how much homogeneity is produced by separating the node at a certain place by comparing the differences between old and new similarities. This equation 11 is used to compute it:

$$Information\ Gain = Left\ Similarity + Right\ Similarity - Similarity\ for\ Root \quad \text{(Equation 11)}$$

- Use the procedure described above to create the required length tree. Playing with the regularization hyperparameter would be used for pruning and regularization.
- Using the Decision Tree created, forecast the residual values.
- The following equation 12 is used to determine the new set of residuals:
    - $New\ Residuals = Old\ Residuals + \rho \sum Predicted\ Residuals$ (Equation 12)
    - Where ρ is the learning rate
- Go back to step 1 and repeat the process foe all the trees.

3.5 Evaluation Criteria

Upon the successful construction of a system model, it undergoes training using the training dataset sourced from the standardized dataset. Subsequently, the model's validation is executed by subjecting it to testing through the utilization of the testing dataset. The efficacy of the system's performance is assessed by Accuracy, Precision, F1 and Sensitivity.

Accuracy:

The ratio of true positives and true negatives to all positive and negative observations is the definition of model accuracy, a performance statistic for machine learning models.

$$Accuracy = \frac{TP+TN}{TP+FN+TN+FP} \quad \text{(Equation 13)}$$

Precision Score:

The percentage of labels that were correctly predicted positively is represented by the model precision score.

$$Precision\ Score = \frac{TP}{TP+FP} \quad \text{(Equation 14)}$$

Recall:

The model's ability to properly forecast the positive out of actual positives is measured by the model recall score.

$$Recall\ Score = \frac{TP}{TP+FN} \quad \text{(Equation 15)}$$

F1:

The F1 score combines precision and recall using their harmonic mean, and maximizing the F1 score implies simultaneously maximizing both precision and recall.

$$F1\ Score = \frac{2*Precision*Sensitivity}{Precision+Sensitivity} \quad \text{(Equation 16)}$$

4. Experiment and Result

4.1 Dataset Overview

In this research dataset is collected form the Kaggle. The dataset originates from the year 1988 and comprises four distinct databases, namely Cleveland, Hungary, Switzerland, and Long Beach V. This dataset encompasses a total of 76 attributes, inclusive of the attribute designed for prediction; however, all publicly reported experiments exclusively employ a subset of 14 attributes. The "target" field pertains to the indication of the presence or absence of heart disease in the patient. Specifically, it is represented as an integer value, where 0 denotes the absence of disease, and 1 signifies the presence of disease. The attribute's dataset specifications are shown in table 1. (kaggle, 2019) (Cenitta, et al., 2022)

Table1: Heart disease attributes

Attribute	Description	Domain of Value	Significance of the attribute in 2 or 3 points so as to designate it as one of the hyper parameters
Age	Age in years	29 to 77	Higher the age, higher the risk of developing coronary artery disease. This happens irrespective of gender, although women tend to be about a decade older when they develop cardiovascular disease compared to men
Sex	Sex	Female (0) Male (1)	Male sex is an independent risk factor for developing heart disease. However, women tend to have poorer outcomes following acute coronary syndromes. Women also have a typical and delayed presentations compared to men
Cp	Chest Pain Type	Typical angina (0) Atypical angina (1) Non-anginal (2) Asymptomatic (3)	Presence of typical angina makes the diagnosis of heart disease much more likely compared to a typical angina. Non-anginal pain makes it less likely. Although uncommon, heart disease can present as silent in elderly patients, diabetics especially with neuropathy etc. however, complete lack of symptoms in a younger, non-diabetic patient usually goes against significant coronary artery disease.
Trestbps	Resting blood sugar	94 to 200 mm Hg	Elevated blood sugar levels esp. fasting blood sugar levels indicates either poor control of sugars in a known diabetic or the presence of diabetes in previous non-diabetic patients. Those with diabetes and heart disease do poorly if their blood sugars are not well controlled with medications.
Chol	Serum cholesterol	126 to 564 mg/dl	Elevated serum cholesterol levels esp. low-density lipoproteins (LDL) are an independent risk factor for heart disease. Control of LDL levels to predefined targets based on the patient's risk profile is one of the main goals of therapy for heart disease.
Fbs	Fasting blood sugar	>120 mg/dl True (1) False (0)	Elevated blood sugar levels esp. fasting blood sugar levels indicates either poor control of sugars in a known diabetic or the presence of diabetes in previous non-diabetic patients. Those with diabetes and heart disease do poorly if their blood sugar is not well controlled with mediations.
Restecg	Resting ECG result	Normal (0) ST-T wave abnormality (1)	Normal resting ECG does not rule out the presence of heart disease. Stress testing like treadmill exercise testing may be required. Presence of LVH can interface with the diagnosis of

		LV hypertrophy (2)	ischemia from both the resting ECG and during treadmill exercise testing.
Thalach	Maximum heart rate achieved	71 to 202	Maximum heart rate achieved during treadmill exercise testing indicates completeness of the test. If the patient achieves a heart rate lesser than this, treadmill test is regarded as inconclusive. If the target heart rate is achieved, then we can go on to interpret the treadmill test further and based on the presence, type and degree of ST-T changes, probability of underlying heart disease is estimated.
Exang	Exercise induced angina	Yes (1) No (0)	Exercise induced angina is an important indicator of significant coronary artery disease. However, it can also be seen in aortic stenosis.
Oldpeak	ST depression induced by exercise relative to rest	0 to 6.2	Larger the ST depression esp. if seen in multiple contiguous ECG leads, higher the likelihood of underlying heart disease.
Slope	Slope of peak exercise ST segment	Upsloping (0) Flat (1) Downslopping (2)	Order of importance from most to least important: down sloping, flat followed by upsloping. Upsloping ST depression is the least important.
Ca	Number of major vessels colored by fluoroscopy	0-3	Coronary angiogram is a diagnostic test used to confirm the presence of coronary artery disease. More the number of vessels affected, worse the clinical outcome for the patient.
Thal	Defeat type	Normal (0) Fixed defect (1) Reversible defect (2)	A reversible defect is specific for significant obstruction in one or more coronary arteries. A fixed defect may be seen in those who have infracted areas indicating previous Myocardial infarction. Normal perfusion study indicates a low chance for having coronary artery disease.
Target	Heart diseases	True (1) False (0)	Target is either heart disease or healthy.

This dataset contains total of 1025 data. Among them 499 data are from healthy people and 526 are from people who had heart disease.

The dataset contains no null value which is verified by the given Table2.

Table 2: Dataset Info

S.N.	Features	Non-Null Count	Dtype
0	Age	1025 non - null	Int64
1	Sex	1025 non - null	Int64
2	Cp	1025 non - null	Int64
3	Trestbps	1025 non - null	Int64
4	Chol	1025 non - null	Int64
5	Fbs	1025 non - null	Int64
6	Restecg	1025 non - null	Int64
7	Thalach	1025 non - null	Int64
8	Exang	1025 non - null	Int64
9	Oldpeak	1025 non - null	Float64
10	Slope	1025 non - null	Int64
11	Ca	1025 non - null	Int64
12	Thal	1025 non - null	Int64
13	target	1025 non - null	Int64

4.2 Experimental Setup and Detail

Python is used to run the experiment on a Windows 10 device with an Intel Core i5, 8th Gen, 1.6GHz CPU, 8GB of RAM, and Anaconda3.0 Scikit-learn 1.0 as the main software platform.

The heart disease dataset from the UCI data repository is used to carry out the experiment. Preprocessing is carried out using Python programming to make the data clean and ready for processing. For this, inter quartile range and min-max normalization are used for outlier discovery and elimination. Following preprocessing, PCA and cumulative explained variance vs number of component analysis are used to determine the number of features.

A majority voting-based ensemble technique is used for the model training, using a variety of machine learning algorithms, including XGBOOST, gradient Boost, Decision Trees, KNNs, and SVMs. At the conclusion of the experiment, accuracy, Recall, precision, and F1 are assessed using testing data models.

4.3 Performance Analysis

4.3.1. Outlier Detection and Removal

In preprocessing phase, outlier detection and removal are performed by using boxplot and inter quartile range formula. Data distribution of a variable against the density distribution is shown in figure 4. And the detailed outlier detection using boxplot is shown in figure 5.

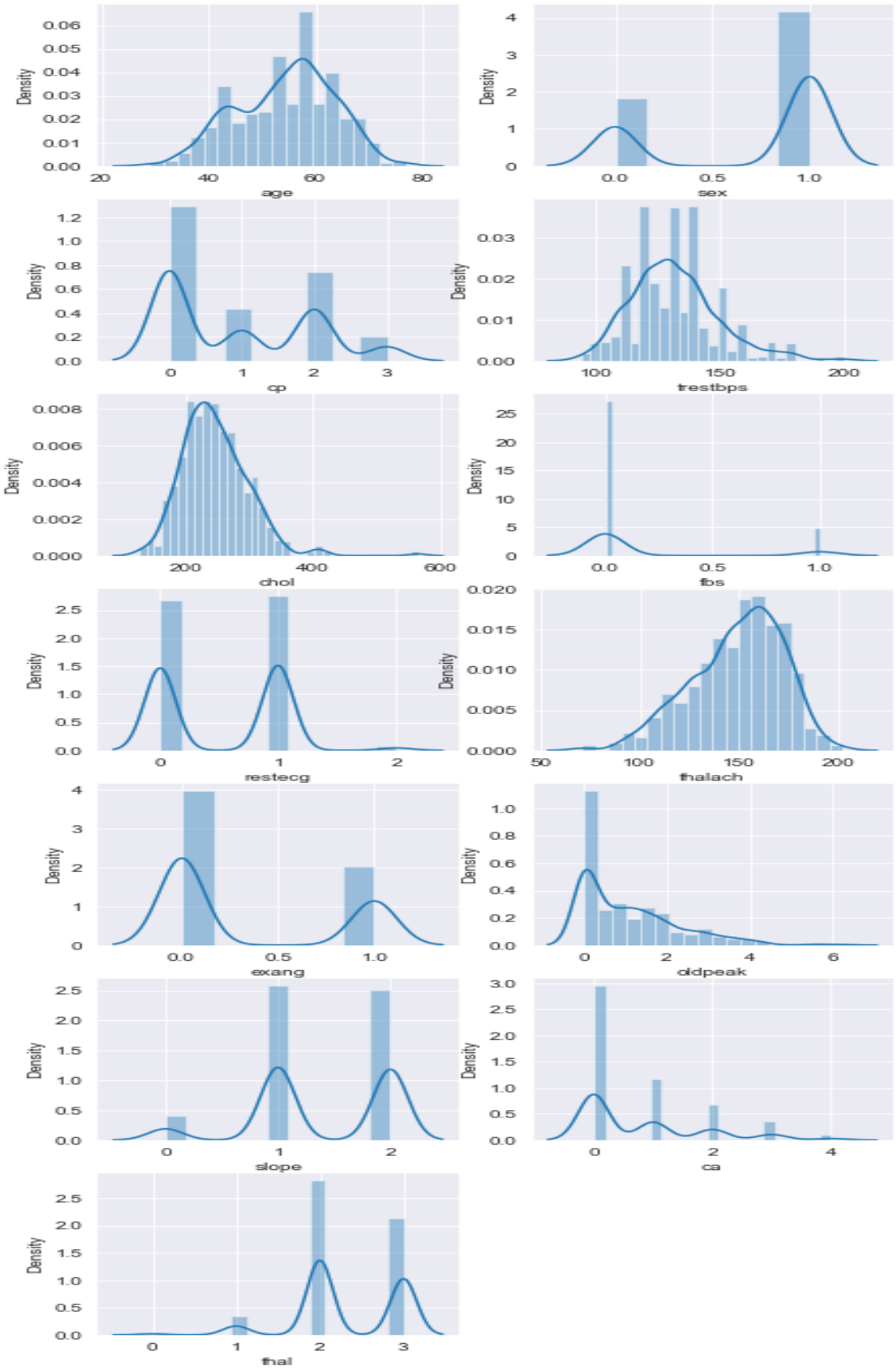

Figure 4: Data distribution of a variable against the density distribution

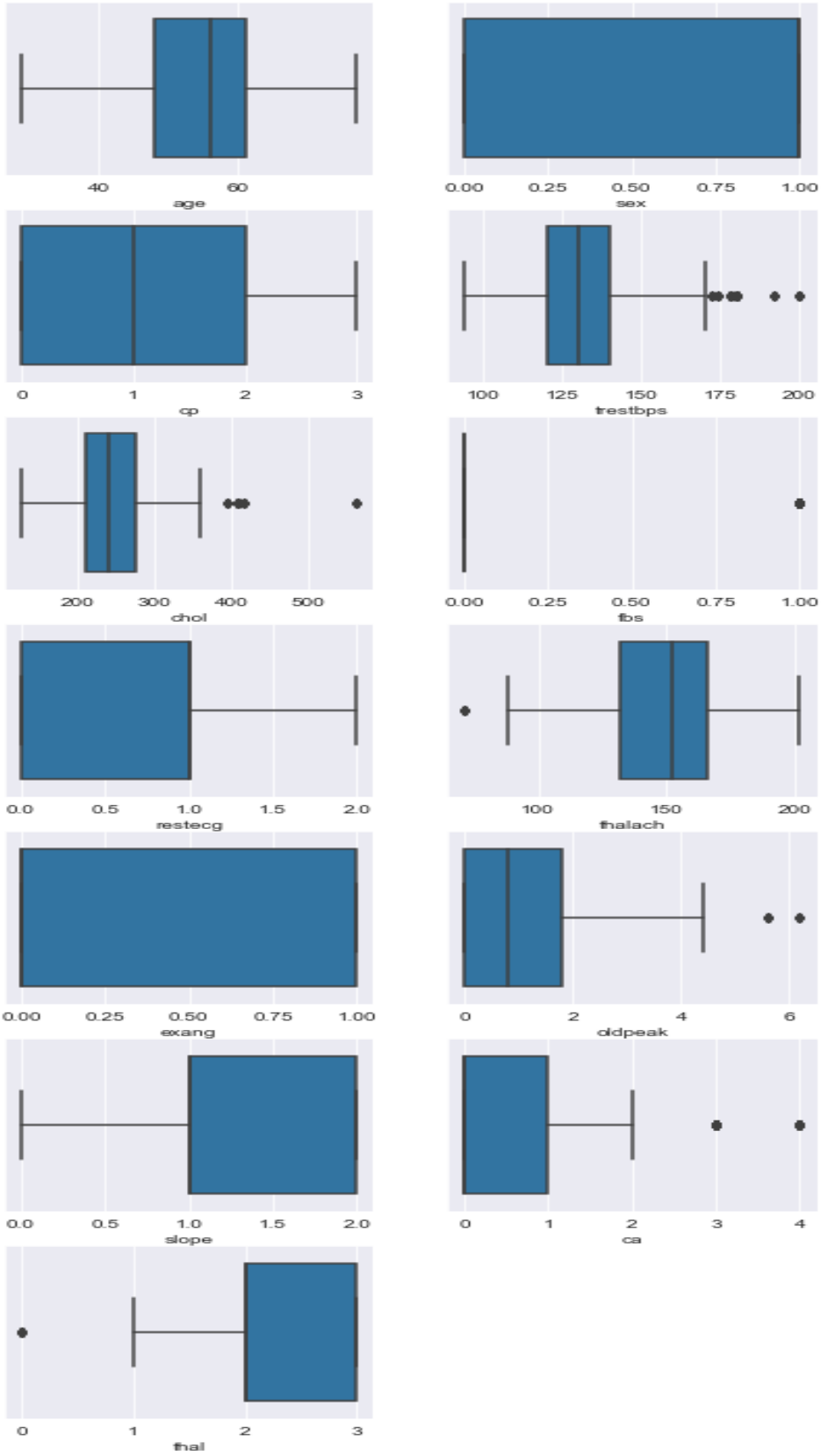

Figure 5: Boxplot of All the features available in dataset

After analyzing the boxplot, it is seen that some of the features are containing the outlier and these outliers are removed by using inter quartile range formula i.e. equation 1,2 and 3. Before and after outlier removal of the features which contains outlier are depicted through the following figure 6,7,8,9,10 and 11.

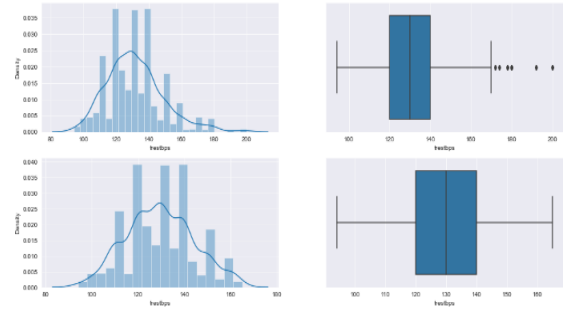

Figure 6: Plot of feature ' trestbps ' before and after outlier removal

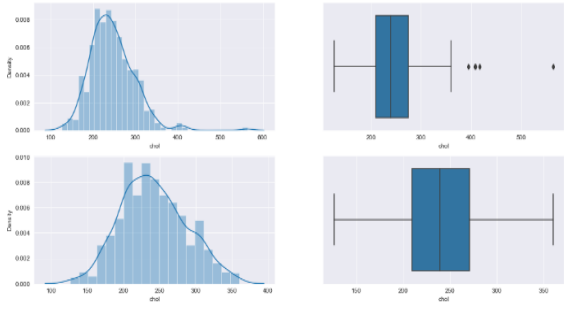

Figure 7: Plot of 'chol' before and after outlier removal

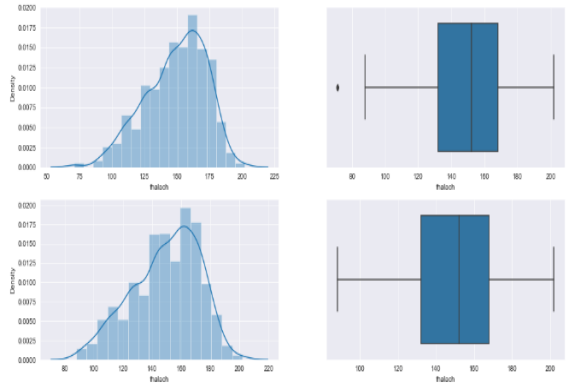

Figure 8: Plot of ' thalach ' before and after outlier removal

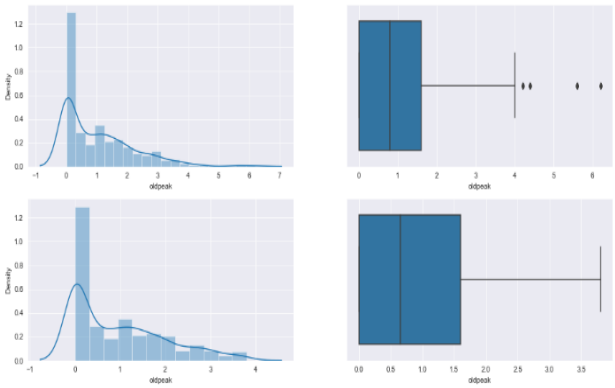

Figure 9: Plot of 'oldpeak 'after and before outlier removal

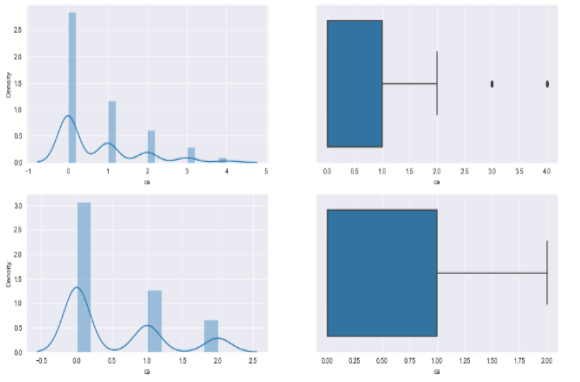

Figure 10: Plot of 'ca 'after and before outlier removal removal

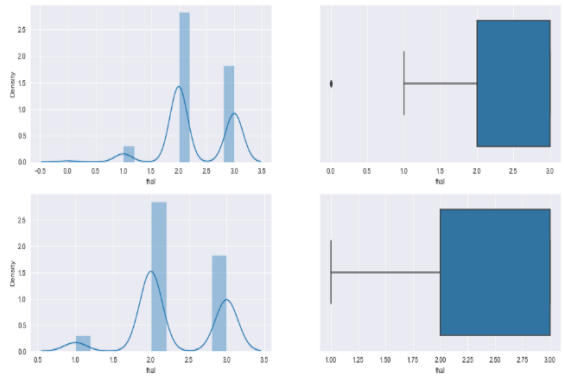

Figure 11: Plot of 'thal 'after and before outlier

After completing the outlier detection and removal the final dataset contains only 862 data samples among which 477 has heart disease and 385 has no heart disease i.e. health. The final dataset distribution is shown in figure 12.

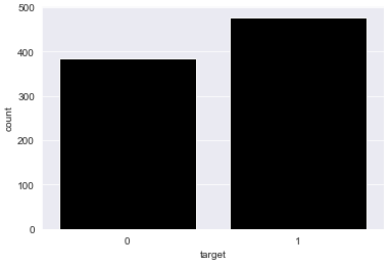

Figure 12: Final Dataset Distribution

4.3.3. Result of the Experiment

After feature selection, 70% of the data are used to tarn the model and remaining 30% are used to test the model. The testing gives the 99.22% of accuracy. Besides this, precision, F1 and sensitivity are also evaluated which are shown in table 3 and bar graph figure 13.

Table 3: Result of the proposed ensemble method

Evaluated Parameter	In %
Accuracy	99.22
Precision	98.68
F1	99.33
Recall	100

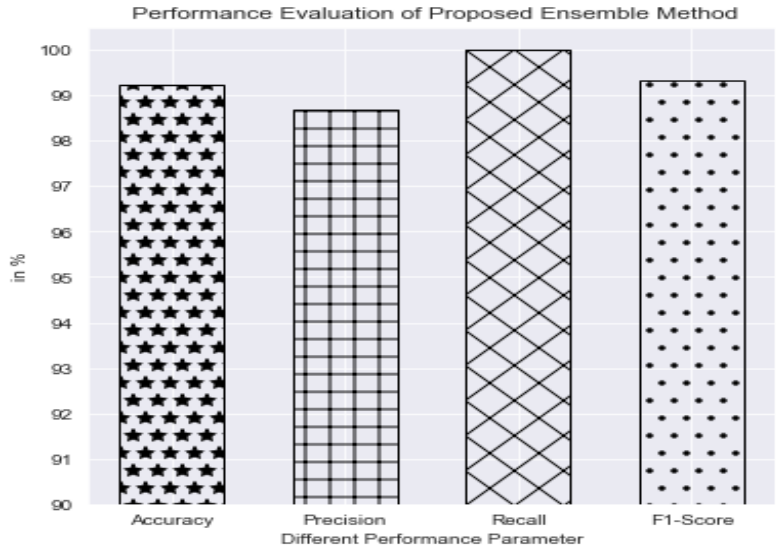

Figure 13: Result of the Ensemble Method

4.3.4. Comparison of the Results

The result obtained through the proposed ensemble method is compared with the three different ensemble method and K Nearest Neighbor (KNN) and Support Vector Machine (SVM). The first ensemble model is considered as Ensemble1 and it employs max voting on Logistic Regression, Multilayer Perceptron and Naïve Bayes. The second ensemble model is considered as Ensemble2 and it uses max voting on stochastic Gradient Descent (SGD), K Nearest Neighbor (KNN), Random Forest (RF), Logistic Regression (LR). The third ensemble model represented as Ensemble3 is also implemented the majority voting ensemble using Random Forest (RF), XGBOOST, Logistic Regression (LR) and K Nearest Neighbor (KNN) algorithms. All the algorithms are implemented on same heart disease dataset obtained from Kaggle for the comparison. The comparison table, table 4 shows the good result of the proposed ensemble method compared to previous method. The comparison is also illustrated in bar graph represented in figure 14,15,16and 17.

Table 4: Comparison of model performance based on Accuracy, Precision, Recall and F1

References	Technique/ Algorithm Used	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
(Raza, 2019)	Ensemble1	82.79	76.70	91.83	83.20
(Atallah & Al-Mousa, 2019)	Ensemble2	90.25	88.74	91.15	90.2
(Rafsun, et al., 2022)	Ensemble3	96.75	97.24	95.91	96.77
(Singh & Kumar, 2020)	K Nearest Neighbor	95.77	95.27	95.91	95.76
(Madhu & Ramesh, 2021)	Support Vector Machine	88.31	82.84	95.23	88.48
Proposed	**MVE (Max Voting Ensemble)**	**99.22**	**98.68**	**100**	**99.33**

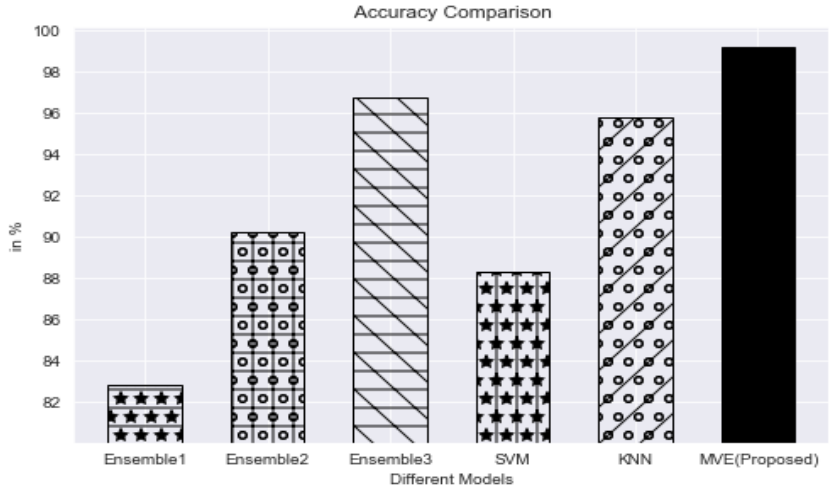

Figure 14: Accuracy Comparison of Different Models

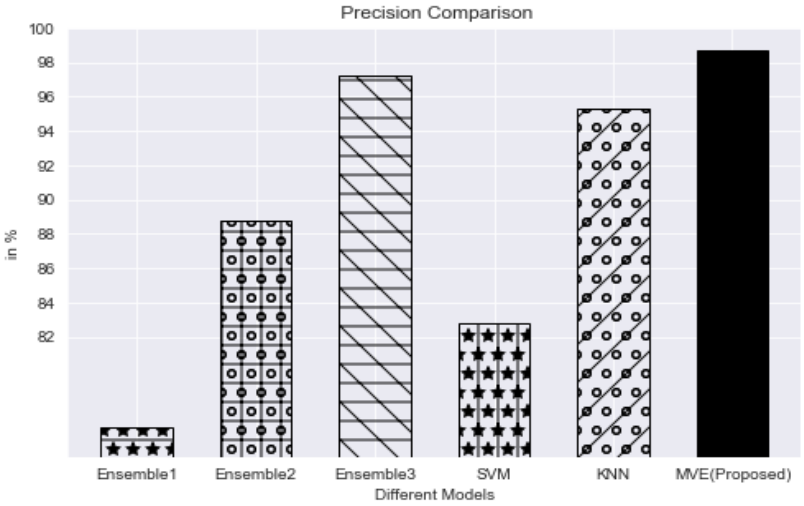

Figure 15: Precision Comparison of Different Models

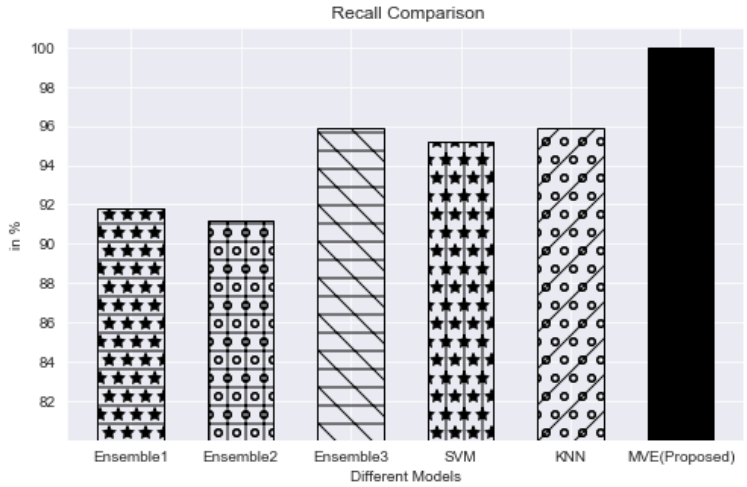

Figure 16: Recall Comparison of Different Models

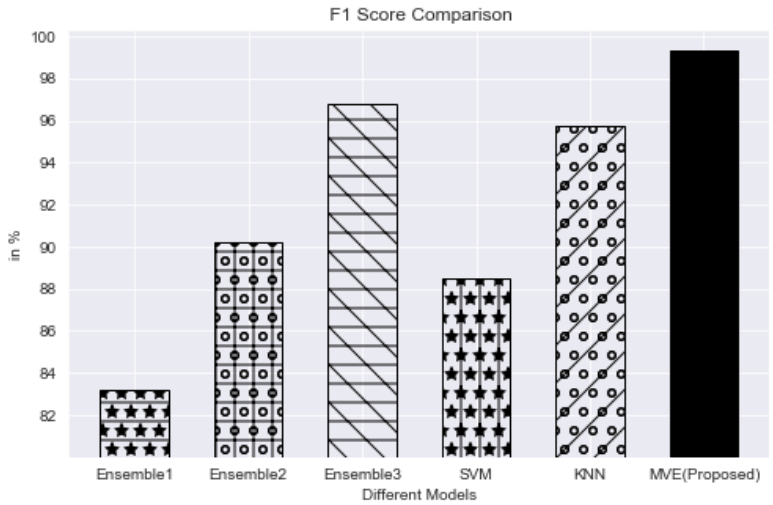

Figure 17: F1 Score Comparison of Different Models

5.Conclusion

Since the human heart is one of the body's most significant organs and heart disease prediction is a major human concern, algorithm accuracy is one of the factors considered when evaluating an algorithm's performance. The dataset utilized for both training and testing purposes affects how accurate machine learning algorithms are. The proposed max voting ensemble method provides the highest accuracy of 99.22%, precision of 98.68%, Recall of 100% and F1 score of 99.33% compared to the previous methods and is an improved models for the heart disease prediction.

In order to reduce the incidence of death cases through increased disease awareness, more machine learning techniques in different dataset will be deployed in the future to analyze cardiac problems more effectively and detect illnesses early.

References

Almazroi, A. A., Aldhahri, E. A., Bashir, S. & Ashfaq, S., 2023. A Clinical Decision Support System for Heart Disease Prediction Using Deep Learning. *IEEE Access,* Volume 11, pp. 61646-61659.

Asif, D., Bibi, M., Arif, M. S. & Mukheimer, A., 2023. Enhancing Heart Disease Prediction through Ensemble Learning Techniques with Hyperparameter Optimization. *Artificial Intelligence Algorithms for Medicine,* 16(6).

Atallah, R. & Al-Mousa, A., 2019. *Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method.* Amman, Jordan, 2nd International Conference on new Trends in Computing Sciences (ICTCS), IEEE, pp. 1-6.

Beunza, J.-J.et al., 2019. Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease). *Journal of Biomedical Informatics,* Volume 97.

Cenitta, D., Arjunan, R. V. & Prema, K., 2022. Ischemic Heart Disease Prediction Using Optimized Squirrel Search Feature Selection Algorithm. *IEEE Access,* Volume 10, pp. 122995-123006.

Chapagain, P., Timalsina, A. & Chitrakar, R., 2022. *Intrusion detection based on PCA with improved K-means.* s.l., Springer Singapore.

Dange, S., Gaikwad, P., Sheral, R. & Shewale, P. D. S. S., 2022. Heart Disease Prediction System using SVM. *International Journal of Innovative Research in Technology,* 8(12).

Hashi, E. K. & Zaman, M. S., 2020. Developing a Hyperparameter Tuning Based Machine Learning Approach of Heart Disease Prediction. *Journal of Applied Science & Process Engineering,* 7(2), pp. 631-647.

Hazra, A. et al., 2017. Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review. *Advances in Computational Sciences and Technology ,* 10(7), pp. 2137-2159.

kaggle, 2019. *Heart Disease Dataset.* [Online]
Available at: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset
[Accessed 02 09 2023].

Kavitha, M. et al., 2021. *Heart Disease Prediction using Hybrid machine Learning Model.* Coimbatore, India, International Conference on Inventive Computation Technologies (ICICT).

Khan, M., 2021. *CSIAS.* [Online]
Available at: https://www.csias.in/explain-the-step-by-step-implementation-of-xgboost-algorithm/
[Accessed 02 09 2023].

Khan, M. A., 2020. An IoT Framework for Heart Disese Prediction based on MDCNN Classifier. *IEEE Access,* Volume 8, pp. 34717-34727.

Li, J. P. et al., 2020. Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare. *IEEE Access,* Volume 8, pp. 107562-107582.

Madhu, H. & Ramesh, D., 2021. Heart Attack Analysis and Prediction using SVM. *International Journal of Computer Applications,* 183(27), pp. 35-39.

Mohan, S., Thirumalai, C. & Srivastava, G., 2019. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *IEEE Access,* Volume 7, pp. 81542-81554.

Qadri, A. M., Raza, A., Munir, K. & S-almutairi, M., 2023. Effective Feature Engineering Technique for Heart Disease Prediction With Machine Learning. *IEEE Access,* Volume 11, pp. 56214-56224.

Rafsun, J., Shanto, M. S. I. & Kabir, M. M. R. M. M. M., 2022. *Heart Disease Prediction and Analysis Using Ensemble Architecture.* Chiangrai, Thailand, IEEE.

Raja, J. B. et al., 2019. Diabetics Prediction using Gradient Boosted Classifier. *International Journal of Engineering and Advanced Technology (IJEAT),* 9(1).

Raza, K., 2019. Improving the Prediction Accuracy of Heart Disease with Enseble Learning and Majority Voting Rule . In: *U-Healthcare Monitoring Systems.* s.l.: Elsevier Inc., pp. 179-196.

Saraswathi, R. V. et al., 2022. *Heart Disease Prediction Using Decision Tree and SVM.* s.l., Proceedings of Second International Conference on Advancesin Computer Engineering and Communication Systems, Algorithms for IntelligentSystems.

Singh, A. & Kumar, R., 2020. *Heart Disease Prediction Using Machine Learning Algorithms.* s.l., 2020 International Conference on Electrical and Electronics Engineering (ICE3).

Sumwiza, K. et al., 2023. Enhanced cardiovascular disease prediction model using random forest algorithm. *Informatics in Medicine Unlocked*, Volume 41.

Taunk, K., De, S., Verma, S. & Swetapadma, A., 2019. *A Brief Review of Nearest Neighbor Algorithm for Learning and Classification.* s.l., International Conference on Intelligent Computing and Control Systems (ICCS).

Verma, S. & Gupta, A., 2021. *Effective Prediction of Heart Disease Using Data Mining and Machine Learning: A Review,*. Coimbatore, India, International Conference on Artificial Intelligence and Smart Systems (ICAIS), pp. 249-253.

Vinutha, H. P., Poornima, B. & Sagar, B. M., 2018. Detection of Outliers Using Interquartile. *Advances in Intelligent Systems and Computing*, Volume 701, p. .

World Health Organization, 2023. *World Health Organization.* [Online]
Available at: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1
[Accessed 02 09 2023].