Ensemble Learning for Heart Disease Diagnosis: A Voting Classifier Approach

Yogesh S, Paneer Thanu Swaroop C, Ruba Soundar K

Abstract*: *Cardiovascular disease remains a serious public health problem internationally, responsible for a considerable number of fatalities. Early and correct detection of cardiovascular illness is crucial for optimal care and control of the condition. In this paper, we present an ensemble learning technique that includes voting classifiers to increase the reliability of cardiovascular disease diagnosis. We obtained a set of data from five cardiology databases, including the Cleveland, Hungary, Switzerland, Long Beach VA, and Statlog (Heart) datasets, which provided us with a total of 1,189 entries. We employed a feature engineering approach to extract relevant features from the dataset, enabling us to acquire vital information to enhance our model's performance. We trained and evaluated several machine learning algorithms, including Random Forests, MLP, K-Nearest Neighbours, Extra Trees, XGBoost, Support Vector Machines, AdaBoost, Decision Trees, Linear Discriminant Analysis, and Gradient Boosting. We then incorporated these models using voting classifiers to produce more reliable and accurate models. Our findings reveal that the proposed ensemble learning process outperforms standalone models and conventional ensemble approaches, obtaining an accuracy rate of 91.4%. Our technique is likely to benefit clinicians in the early diagnosis of heart problems and improve patient outcomes. This work has significant implications for the field of cardiology, suggesting that machine learning approaches may enhance both the reliability and accuracy of heart disease identification. The recommended ensemble learning technique can be applied in hospitals to improve patient care and ultimately mitigate the global impact of cardiovascular disease. Further study is needed to investigate the applications of predictive modelling in cardiology and other medical fields.

Keywords: Cardiovascular Disease; Heart Disease Diagnosis; Ensemble Learning; Voting Classifier; Machine Learning; Feature Engineering; Cleveland Dataset; Hungary Dataset; Switzerland Dataset; Long Beach VA Dataset; Statlog (Heart) Dataset

I. INTRODUCTION

Heart disease is a severe global medical issue, creating a significant amount of mortality each year. According to data from the World Health Organization (WHO), coronary artery disease accounts for around 17.9 million deaths per year [1], making it the top cause of death internationally. Among the various kinds of cardiovascular conditions, heart disease ranks as one of the most prevalent and life-threatening. Heart disease involves a variety of conditions that impair the way the hearts operate, including clogged coronary arteries, cardiac arrhythmia, and heart failure [2]. Early identification and accurate diagnosis of cardiovascular disorders are crucial for effective management and treatment of the condition. The usual techniques of assessment, including electrocardiography (ECG) and echocardiography, have limitations regarding accuracy and reliability. There exists an urge for improved and more successful methods of identifying cardiac disease [3]. Machine learning approaches have emerged as potential tools for enhancing the accuracy and dependability of cardiac disease detection. Machine learning algorithms may learn from data and detect patterns that are not observable by humans. These algorithms may be trained on enormous amounts of patient data and medical pictures to detect early indications of heart failure and forecast the chance of acquiring cardiac disease in future years [4]. The implementation of machine learning methods in medicine has already demonstrated promising outcomes. Several studies have claimed promising findings in the early identification and diagnosis of cardiac conditions using predictive machine learning algorithms [5]. Yet, the precision and dependability of these algorithms may be improved by employing ensemble learning approaches. Ensemble learning is a machine learning method that combines multiple models to enhance the overall system's efficiency and accuracy. The concept driving ensemble learning is to construct a set of diverse approaches that can complement each other's capabilities and address their shortcomings. By merging the predictions of different models, the precision as well as the dependability of the system may be increased [6].

In this paper, we present an ensemble learning technique for heart disease detection that incorporates voting classifiers. We utilized four cardiology datasets from the Machine Learning Repository at UCI - the Cleveland, Hungary, Switzerland, and Long Beach VA data sets [7] - to construct a dataset of 1189 entries. We employed a feature engineering strategy to extract significant properties from the dataset, enabling us to acquire critical information that enhances the performance of our model. We trained and tested many machine learning methods, such as Random Forests [8], MLP, K-Nearest Neighbors, Extra Trees, XGBoost, Support Vector Machines, AdaBoost, Decision Trees, Linear Discriminant Analysis, and Gradient Boosting, to develop our base models.

Manuscript received on 07 August 2023 | Revised Manuscript received on 10 October 2023 | Manuscript Accepted on 15 November 2023 | Manuscript published on 30 November 2023.

*Correspondence Author(s)

Yogesh S*, B.E, Department of Computer Science and Engineering, Mepco Schlenk Engineering College, Sivakasi (Tamil Nadu), India. Email: yogeshselvarajan@gmail.com, ORCID ID: 0009-0008-1568-486X

Paneer Thanu Swaroop C, B.E, Department of Computer Science and Engineering, Mepco Schlenk Engineering College, Sivakasi (Tamil Nadu), India. Email: tcenthilraj@gmail.com

Ruba Soundar K, Associate Professor (Sr. Grade), Department of Computer Science and Engineering, Mepco Schlenk Engineering College, Sivakasi (Tamil Nadu), India. Email: rubasoundar@gmail.com, ORCID ID: 0000-0003-1300-6519

These models were then combined through voting classifiers to build more reliable and accurate models.

To evaluate the success of our proposed ensemble learning method for detecting heart disease, we applied various performance criteria. We tested our models using measures such as Accuracy, Precision, Sensitivity, Specificity, F1 Score, ROC, and Matthews Correlational Coefficient. These measures enabled us to assess the accuracy, dependability, and overall success of our method. The dataset used in our investigation comprises a total of 1,189 records and 12 characteristics. The characteristics included patient information such as age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, rest ECG, maximum heart rate reached, exercise-induced angina, ST depression, ST slope, and goal. The target characteristic, which was our primary emphasis, had binary values of 0 and 1, where 0 indicated the absence of heart illness and 1 indicated the existence of heart disease [9].

Overall, our study makes a significant contribution to the field of predictive modelling in medicine, particularly in the domain of heart disease identification. Our technique has the potential to enhance the accuracy and reliability of existing approaches for diagnosing cardiovascular disease, potentially contributing to more efficient treatment and management of this life-threatening illness.

II. RELATED WORKS

Up to today, several studies have been conducted on the early diagnosis of coronary artery disease and heart failure. They have utilised several machine learning prediction methods and achieved impressive performance. This section includes a complete literature review of research works in the domain of heart disease diagnostics underpinned by machine learning techniques:

In [10], the authors present an improved use of machine learning techniques that predicts heart disease probability. Their strategy involves randomly splitting the data set and modelling each partition using both regression and classification tree methods. A homogeneous ensemble is built from multiple models using an accuracy-based weighted ageing classifier ensemble. This technique obtained 93% and 91% classification accuracies on the Cleveland and Framingham datasets, respectively, exceeding existing machine learning algorithms and related efforts. The findings demonstrate the effectiveness of the proposed ensemble learning technique for predicting heart disease risk.

In [11], the authors employed multiple machine learning techniques including deep learning to examine the UCI Machine Learning Heart Disease dataset. They handled extraneous characteristics utilizing Isolation Forest and normalized the findings. The research yielded encouraging findings, which were confirmed by applying accuracy and confusion matrices. Additionally, the research examined the possible coupling of this study with multimedia technologies on mobile devices. The deep learning method yielded an accuracy rate of 94.2%.

In [12], the authors created an ensemble model for heart disease prediction utilizing the University of California, Irvine ML Repository. The meta-algorithm demonstrated a steadier performance than baseline machine learning algorithms, resulting in higher prediction accuracy and more dependable analytical output. The Bagging Ensemble classifier was determined to have the best prediction probability score and was recommended for deployment. The authors provided a cost-effective and user-friendly graphical user interface for the ensemble heart disease prediction model, which is scalable and extendable. These data suggest that the ensemble model is a viable alternative for predicting heart disease.

In [13], the authors developed a deep stacking ensemble model for predicting heart disease, which incorporates two pre-trained blended deep learning models (CNN-LSTM and CNN-GRU) with SVM acting as the meta-learner model. Recursive Feature Elimination (RFE) was employed for feature selection optimisation, and the proposed model was compared with five different machine learning models. The suggested model achieved its maximum performance by utilising the entire feature set.

In [14], the authors suggest a Stacking CVClassifier ensemble model built of Logistic Regression, K Nearest Neighbors, and Naïve Baye's classifiers as base learners to boost the diagnostic performance of these classifiers for predicting heart disease. They preprocessed the information from the Cleveland Clinic Foundation, utilising various data extraction and feature selection methods to enhance the accuracy of illness prediction. The suggested ensemble model achieved a prediction accuracy of 90.0% on the test data, outperforming the component models. The authors argue that such models may aid healthcare practitioners in developing medical diagnostic systems, particularly in disadvantaged regions, and offer a non-invasive method for diagnosing illnesses.

In [15], the authors present a novel ensemble model, named RSS-KNN, for early prediction of heart disease. The framework combines the stochastic subspace and K-nearest neighbour (KNN) techniques with an isolation-based outlier elimination approach and an enhanced squirrel optimiser. The suggested technique captures key information through RSS and then feeds it into KNN for accurate classification. The simulation results on the UCI ML dataset demonstrate that the ensemble model suggested attains an accuracy of 98.56% with features and 98.10% specificity, exceeding the accuracy of current modern facility classifiers.

In [16], the authors apply machine learning approaches to identify heart problems in people. The work combines both heterogeneous and homogeneous ensemble classifiers, leveraging the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance and noise. The suggested technique incorporates two steps: SMOTE preprocessing and classification utilizing Naive Bayes (NB), decision tree (DT), and their ensembles. The findings reveal that the AdaBoost-Random Forest predictor achieves an accuracy of 95.47% in the early detection of heart disease.

In [17], the authors present a heart disease prediction model that incorporates body signals and an algorithm centered upon the integrated datasets and train-test split approach. They compare their findings with prior efforts and employ accuracy, precision, recall, F1-score, and ROC-AUC curves as performance measures for various classifiers.

The Random Forest Classifier achieves 100% performance on the combined heart disease datasets, whereas the Decision Tree, Random Forest, and Gradient Boosting Classifiers yield ROC-AUC scores of 99-100%. They also employ five-fold cross-validation and the Stacking CV Classifier to enhance the performance of the machine learning algorithms. The authors discovered that the RFC classification algorithm's accuracy is excellent, and their proposed approach is efficient and accurate for predicting heart disease. Overall, the studies demonstrate the effectiveness of ensemble learning and feature selection in enhancing heart disease prediction accuracy. The results of the research can help develop more accurate and robust models for predicting heart disease, which can aid in early diagnosis and treatment of the condition.

III. COMPUTATIONAL INTELLIGENCE TECHNIQUES

Computational Intelligence Techniques refer to the set of methods and algorithms that enable computers to learn from data, adapt to changing environments, and perform tasks that traditionally require human intelligence. These techniques encompass a wide range of subfields, including machine learning, artificial intelligence, and data mining. Machine learning algorithms are a key component of Computational Intelligence Techniques. These algorithms learn from data to improve performance on a specific task. In this section, we will discuss several popular machine learning algorithms that have been used to develop the ensemble model, including Random Forests, MLP, K-Nearest Neighbours, Extra Trees, XGBoost, Support Vector Machines, AdaBoost, Decision Trees and Gradient Boosting [18].

A. Random Forests

Random Forest is a powerful ensemble learning technique used for both classification and regression tasks. It builds a forest of decision trees on random subsets of data and features to mitigate the risk of overfitting and enhance the model's accuracy. In this paper, we utilized the Random Forest Classifier to implement the Random Forest algorithm with two different sets of parameters. The first set included the criterion='entropy' parameter, which measures the quality of splits, and n_estimators=100, which specifies the number of trees in the forest. The second set of parameters was criterion='gini' and n_estimators=100. This allowed us to compare the performance of Random Forest with different criteria for measuring the quality of splits.

B. Multi-Layer Perceptron (MLP)

The Multi-Layer Perceptron (MLP) is a type of artificial neural network that is used for classification tasks. It has an output layer with numerous neurons connected by weighted connections, one or more hidden layers, and an input layer. The MLP classifier is a supervised learning algorithm that adjusts the weights between neurons using a backpropagation algorithm to minimise the error between the predicted and actual outputs. We used the MLPClassifier() as one of the base models in this paper.

C. K-Nearest Neighbour (KNN)

K-Nearest Neighbour (KNN) is a classification algorithm commonly used in data mining, particularly for medical databases. It works by measuring the similarity between a new case and other cases in the dataset. The algorithm calculates the distance between the new case and every other case in the model and then assigns the new case to the output that contains the closest neighbours. The KNN algorithm has two steps:

1) finding the K training instances that are closest to the unidentified instance.
2) selecting the most frequently occurring classifications for these K instances.

In this paper, we have used the KNeighborsClassifier() to implement the KNN algorithm with K=9, which means that the algorithm considers the nine nearest neighbours to the new instance for classification.

D. Extra Trees

Extra Trees is an ensemble learning method that uses a combination of multiple decision trees to generate a forest. It constructs many random decision trees and averages their output to produce the result. This approach helps reduce overfitting and improve the model's accuracy. In Python, the ExtraTreesClassifier () function is used to implement the Extra Trees algorithm for classification problems. The parameter n_estimators specifies the number of trees to be used in the forest. In this paper, we used three Extra Trees Classifier models as base models for our ensemble model, with n_estimators set to 100, 500, and 1000, respectively.

E. XGBoost

XGBoost, short for eXtreme Gradient Boosting, is a gradient boosting algorithm used for supervised learning tasks. It is designed to enhance the performance of traditional gradient boosting algorithms by incorporating a regularisation term into the loss function, which helps prevent overfitting. XGBoost builds multiple decision trees iteratively, with each tree attempting to correct the errors of the previous tree. The algorithm assigns weights to each instance in the data, with the weights of misclassified instances increasing in subsequent iterations. In this paper, we utilized the XGBClassifier() to create five models. The first model was built without any additional parameters, while the second model had n_estimators set to 2000. The third model was created with n_estimators set to 500, the fourth with n_estimators set to 100, and the fifth with n_estimators set to 1000. These models served as the basis for our ensemble learning approach, which aimed to enhance the classification performance on our medical dataset.

F. Support Vector Machines (SVM)

The Support Vector Machine (SVM) has demonstrated excellent performance in the medical sector for disease prediction in recent years. SVM is a supervised learning technique designed for regression and classification tasks, with the primary goal of minimizing generalization errors. SVM works by classifying data into two classes over a hyperplane. It is highly effective in high-dimensional spaces, even when the number of dimensions exceeds the number of samples. Mathematically, SVM is represented as follows.

Retrieval Number: 100.1/ijese.J255509111023
DOI: 10.35940/ijese.J2555.11111223
Journal Website: www.ijese.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

$$\text{If } Y_i = +1;\ wx_i + b \geq 1 \quad (1)$$

$$\text{If } Y_i = -1;\ wx_i + b \leq 1 \quad (2)$$

$$\text{For all } i;\ y_i\,(w_i + b) \geq 1 \quad (3)$$

In the equation, x is a vector point, and w is a weight and a vector. So, the data in Equation (1) should always be greater than zero, and the data in Equation (2) should always be less than zero. Among all possible hyperplanes, SVM selects the one where the distance of the hyperplane is as significant as feasible. In this paper, we have used the SVC() with the parameters kernel='linear', gamma='auto', and probability=True. The Support Vector Classifier (SVC) is a type of SVM algorithm used for classification tasks. The kernel parameter specifies the type of kernel function to be used for the decision function, while the gamma parameter controls the width of the kernel function. The probability parameter is set to True to enable probability estimates for each class.

G. AdaBoost

AdaBoost is an ensemble learning method used for classification problems. It combines multiple weak classifiers to create a strong classifier. The weak classifiers are typically decision trees with a depth of 1 and are trained on subsets of

the data. During training, the algorithm assigns weights to the samples in the dataset and trains a new weak classifier in each iteration. The final classifier is a weighted sum of the weak classifiers, where each classifier is assigned a weight based on its accuracy. In this paper, we used the AdaBoost Classifier function without any parameters.

H. Decision Tree

Decision trees are powerful and popular tools for predicting and classifying medical data. A decision tree is a classifier expressed as a recursive part of the information space, depending on the values of the attributes. Each internal node splits the instance space into two or more subspaces according to the specific function of the input attribute values. Each leaf is allocated to one class that represents the most appropriate value. There are many types of algorithms for DT. In this paper, we used the Decision Tree Classifier as one of the base models in our ensemble model. We did not specify any additional parameters for this model, meaning that it used the default parameters to build the decision tree.

I. Gradient Boosting:

Gradient Boosting is an ensemble learning technique used for classification and regression problems. It sequentially builds trees by giving more weight to the misclassified instances in the previous tree. The Gradient Boosting Classifier in Python uses the gradient boosting algorithm for classification problems, with n_estimators and max_features as essential parameters. In our research, we used the Gradient Boosting Classifier with n_estimators=100 and max_features='sqrt' as one of the base models for our ensemble model.

IV. METHODS AND MATERIALS

In this study, to forecast the coronary artery heart disease, we use the following procedures, (1) Data collection, (2) Exploratory Data Exploration (EDA), (3) Data preprocessing, (4) Base Models Evaluation, (5) Feature Selection, and (6) Voting Ensembling. The following sections provide in-depth coverage of all the processes.

A. Data Collection

The data collection method is an integral part of every research study, as it ensures the quality and confidence of the results. In this work, all five datasets were collected using the heart disease dataset available via the UCI Machine Learning Repository. The Heart Disease dataset is a comprehensive database commonly used in studies related to heart disease. The dataset utilized in this research is generated by combining five significant heart disease datasets, including Cleveland, Hungarian, Switzerland, Long Beach VA, and Statlog (Heart) Data Set, over 12 standard features. This unification procedure has resulted in a single dataset that is the greatest heart disease dataset available for research purposes. By merging data from multiple sources, the dataset provides a more comprehensive picture of cardiac disease, including its causes, risk factors, and potential therapies. The Heart Disease dataset encompasses both male and female patients, providing a diverse sample for heart disease research. With an aggregate total of 1189 samples divided into 11 categories, the dataset provides a plethora of information for academics to evaluate. The 11 factors presumably include the patient's clinical assessment information, which may provide insights into the development and progression of cardiac disease. The target categorisation, which refers to the presence or absence of cardiac disease, is the 12th characteristic, making it the subject of this investigation. The availability of such a vast and diversified dataset is a huge benefit for researchers investigating cardiac disease. The scale and variety of the dataset enable more accurate and dependable discoveries, which may eventually lead to improved diagnosis, prevention, and treatment of cardiac disease. More information regarding the dataset can be found in Table 1. Overall, the Heart Disease dataset used in this study is a valuable resource for researchers investigating heart disease and its associated illnesses.

B. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial component in every data-driven research endeavour. It involves assessing and summarising data to gain insights and knowledge about underlying patterns and relationships between variables. EDA tools, such as visualisations and statistical summaries, can help identify outliers, missing values, and irregularities in the data. By employing EDA, researchers can gain a deeper understanding of the data and identify any issues that may need to be addressed before further investigation. The goal of doing EDA on the dataset is to acquire insights into the distribution and relationships between variables. The information comprises numerous demographic, medical, and clinical variables, and examining these features may help reveal potential risk factors and inform preventive strategies. Additionally, EDA may help identify any missing or erroneous data points that could compromise the accuracy of subsequent research.

Table 1: Description of the Heart Disease Dataset

S. No	Attribute	Datatype	Value
(i)	age	Numeric	[28 – 77]
(ii)	sex	Binary	1 = male, 0= female;
(iii)	chest pain type	Nominal	1 indicates typical angina, 2 indicates atypical angina, 3 indicates non-anginal pain, and 4 indicates no symptoms;
(iv)	resting BP s	Numeric	[94–200]
(v)	cholesterol	Numeric	[126–564]
(vi)	fasting blood sugar	Binary	1,0; > 120 mg/dl
(vii)	resting ecg	Numeric	According to Estes' criteria, 0 represents normal, 1 represents an abnormal ST-T wave, and 2 represents either definite or likely left ventricular hypertrophy.
(viii)	max heart rate	Nominal	[71–202]
(ix)	exercise angina	Binary	1 = yes; 0 = no
(x)	oldpeak	Numeric	1 = yes; 0 = no
(xi)	ST slope	Nominal	1 = upsloping; 2 = flat; 3 = downsloping;
(xii)	target	Binary	1 = heart disease; 0 = Normal;

Figure 1 depicts the correlation coefficients among every pair of features in the dataset. The graphic was built using Seaborn's heatmap() function with the "PuBuGn" colour map and illustrates the correlation values between each feature. Purple denotes a weaker correlation, whereas green represents a stronger correlation.

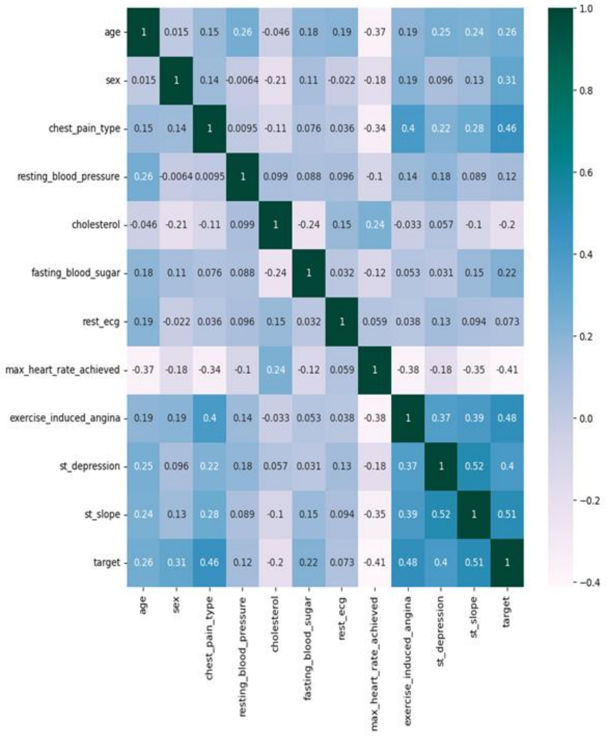

Figure 1: Correlation Heatmap of the Dataset

Figure 2 illustrates the distribution of patients with normal and heart disease in this dataset. It displays the count of the patients on the x-axis and both categories on the y-axis. The green bars represent the normal patients, whereas the red column depicts the patients with heart disease. The count for every category is presented at the end of the column.

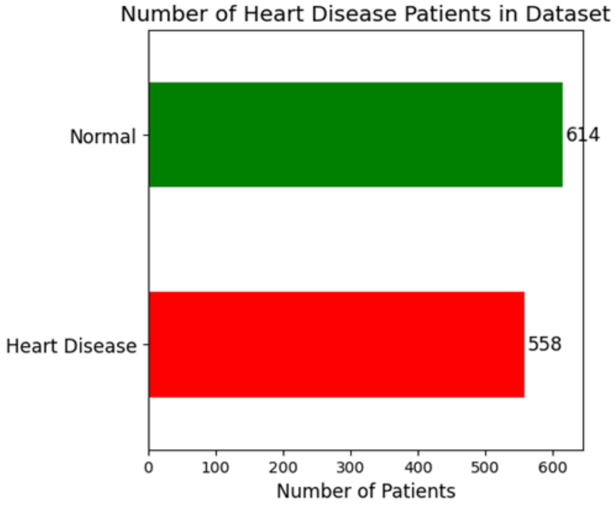

Figure 2: Number of Normal and Heart Disease Patients

Figure 3 shows the distribution of gender in the dataset, which was investigated. The findings reveal that 76.4% of the people were male, while 23.6% were female. This observation suggests that the dataset may be biased towards a male population, which could potentially affect any gender-related research or conclusions drawn from the data.

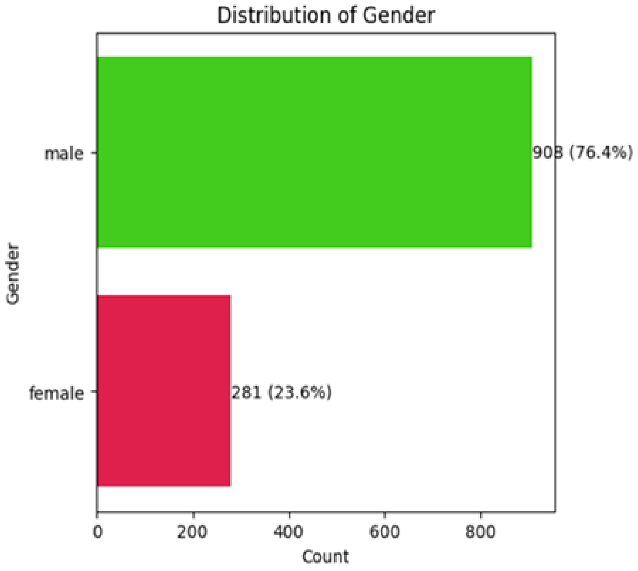

Figure 3: Count of Gender

In this research initiative, we conducted an in-depth investigation of the dataset, yielding vital insights into the distribution and interactions between the variables. We have identified potential risk factors that may contribute to heart disease and emphasised the importance of correcting missing or incorrect data. Furthermore, we have identified a gender bias in the dataset, which may have implications for gender-related research or conclusions drawn from the data. With these insights in mind, we will go to the data processing step, where we will prepare the dataset for future analysis.

C. Data Preprocessing

Data preprocessing is a vital stage in every machine learning effort. The performance of classification problems is heavily reliant on the quality and standard of the dataset used for training and testing the models. One common problem that may arise in datasets is the presence of missing values, which can compromise the accuracy and reliability of the conclusions drawn. Therefore, it is vital to detect and manage missing information effectively before continuing with any analysis.

There are numerous techniques for addressing missing values, including altogether disregarding them, replacing them with a numeric value such as zero, replacing them with the most common value for that property, or replacing them with the mean value for that attribute. It is essential to note that the dataset used for this research work contains no missing values. As a result, handling missing values is not necessary, and we can proceed to the following stage of data preparation. To make the categorical variables more intelligible, we have applied label encoding to the 'chest_pain_type', 'rest_ecg' and 'st_slope' columns. During the exploratory data analysis phase, it was discovered that several numerical characteristics in the dataset contained outliers. Specifically, the age, resting blood pressure, cholesterol, and maximum heart rate reached had outliers that might impair the accuracy and dependability of the machine learning models. Therefore, it is vital to detect and treat outliers effectively before continuing with any study. Various approaches, such as z-score, interquartile range, or Tukey's fences, may be used to detect and treat outliers. In this work, we employed the z-score approach to detect and eliminate outliers from the dataset, ensuring the correctness and reliability of the findings obtained from the machine learning models. To address outliers in the dataset, a threshold value of 3 was specified, which was used to detect data points that were more than three standard deviations away from the mean. This was performed using the np.where() function to find the indices of all data points that exceed the threshold value, and then filtering out those points from the dataset using the all() method with the axis=1 option. After eliminating outliers from the sample, the size of the dataset dropped from (1190, 12) to (1172, 12). This indicates that a total of 18 data points were identified as outliers and subsequently removed from the dataset. After outlier elimination, categorical variables in the dataset were encoded using one-hot encoding. The generated dataset consisted of a total of 15 columns, comprising five numerical variables and 10 categorical features. To analyze the performance of our model, we divided the data into training and testing datasets. In this paper, the data is divided into two sets: one for training and one for testing. The stratify option is set to y, which guarantees that the datasets used for training and testing have the same percentage of target variable classes as the original dataset. The test_size option is set to 0.3, which implies that 30% of the data is utilized for testing, while the remaining 70% is used for training. The shuffle option is set to True, indicating that the data undergoes shuffling before splitting. The random_state parameter is set to 5, ensuring that the findings are repeatable. To ensure that all numerical features are on the same scale, we normalised the dataset using the MinMaxScaler from the scikit-learn module. The scaler stabilises data within the range of 0 to 1. The age, resting blood pressure, cholesterol, maximum heart rate reached, and depressive characteristics were all adjusted using this procedure. This normalising method helps ensure that no single characteristic dominates the others and that all features contribute equally to the study. In summary, the data preparation phase encompasses several stages, including resolving missing values, eliminating outliers, and encoding categorical variables. After completing all the preparation stages, we ended up with a total of 820 data points in the training set and 352 data points in the test set, both of which contained 15 columns. These datasets will be utilised to train and assess our machine learning model in the next part of the research work.

D. Base Model Evaluation

In this work, several machine learning approaches were employed to forecast the incidence of heart disease. The implementation specifics of these models were described in Section III. To further enhance prediction accuracy, an ensemble model utilising the Voting Classifier approach was constructed. The Voting Classifier is a straightforward approach to aggregating predictions from multiple models, thereby enhancing the overall performance of the model. The ensemble model was trained on the training data and deployed to predict the labels of the test data. The implementation specifics of the Voting Classifier model are as follows: initially, a VotingClassifier object was constructed with estimators comprising all the individual models. The 'hard' voting approach was utilized, which means the projected class labels of each model were used to create the final forecast. The ensemble model was then trained on the training data and used to predict the labels of the test data.

After building a voting classification model that combines the results of the individual models, the model is evaluated using a confusion matrix. The confusion matrix is used to illustrate the number of accurate and incorrect predictions produced by the model. The actual negative (TN), false negative (FN), true positive (TP), and false positive (FP) values are computed from the confusion matrix. This assessment helps determine the model's usefulness in predicting heart disease outcomes. Figure 4 illustrates the confusion matrix associated with the voting classifier model. The TN value is 150, the FP value is 22, the FN value is 14, and the TP value is 170.

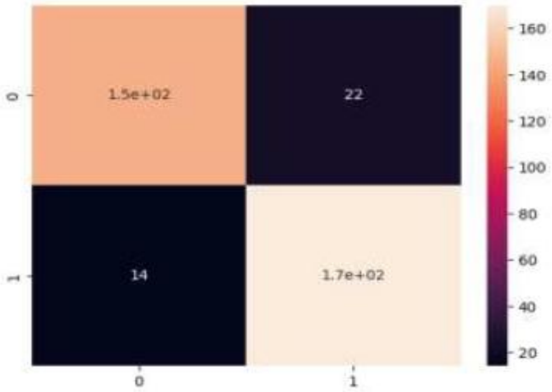

Figure 4 : Confusion Matrix for the Ensemble Model on the Test Set

E. Feature Selection

The feature selection process involves identifying the most relevant features from a dataset that contribute the most towards a model's predictive performance. It helps to remove redundant or irrelevant features, which can lead to overfitting and reduced model performance. Based on the Random Forest classifier, we have performed feature selection by selecting the most essential features. The feature importances have been plotted to identify the most critical features, and a threshold of 0.05 has been used to select the most significant ones. The selected features have been used to create new training and testing sets, and the other less essential features have been discarded. Figure 5 shows the feature importance plot generated by the Random Forest classifier. The plot displays the relative importance of each feature, with higher bars indicating greater importance.

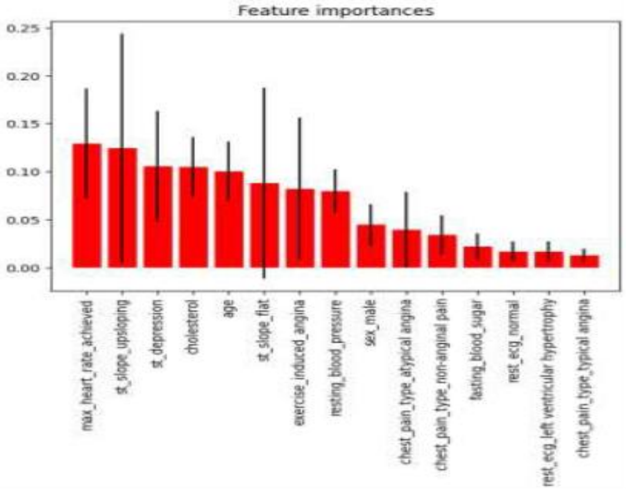

Figure 5 : Feature Importance Plot Generated using Random Forest Classifier.

The feature importance plot generated using the Random Forest classifier shows that the most important features for predicting heart disease are the maximum heart rate achieved, ST slope during upsloping, ST depression, cholesterol level, age, ST slope during flat, presence of exercise-induced angina, resting blood pressure, and male sex. The least relevant characteristics include chest pain kind (atypical angina, non-anginal discomfort, typical angina), fasting blood sugar level, and resting ECG findings (normal, left ventricular hypertrophy). Next, these essential attributes will be utilised to develop a new ensemble voting classifier, which is expected to improve the accuracy of the initial ensemble model substantially.

F. Voting Ensembling

In this step, we defined a new voting classifier using soft voting and employed the same machine learning models described in Section III as estimators. The main difference is that the base voting classifier model has undergone feature engineering, where only the best features selected in the previous step are fed into the model. The voting classifier is then trained using the training set and evaluated using the test set. After training and assessing the voting classifier using the given features, we constructed a confusion matrix. We estimated the true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). We then showed the confusion matrix using a heatmap with annotations to illustrate the number of correct and wrong predictions for each class. Figure 6 illustrates the confusion matrix of the feature-engineered Voting Classifier on the test set. In this case, the confusion matrix has TN = 98, FP = 13, FN = 7, and TP = 120.

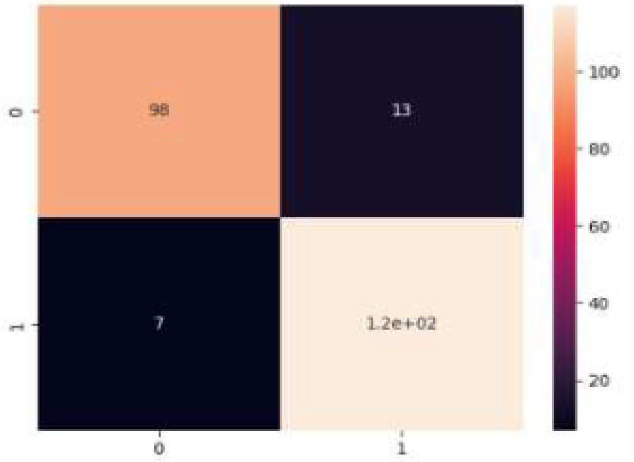

Figure 6 : Confusion Matrix of the Feature-Engineered Voting Classifier.

Figure 7 illustrates the comprehensive system design, providing a concise overview of the proposed ensemble learning approach for heart disease diagnosis. The system is designed to address the critical need for accurate and reliable detection of cardiovascular disease.

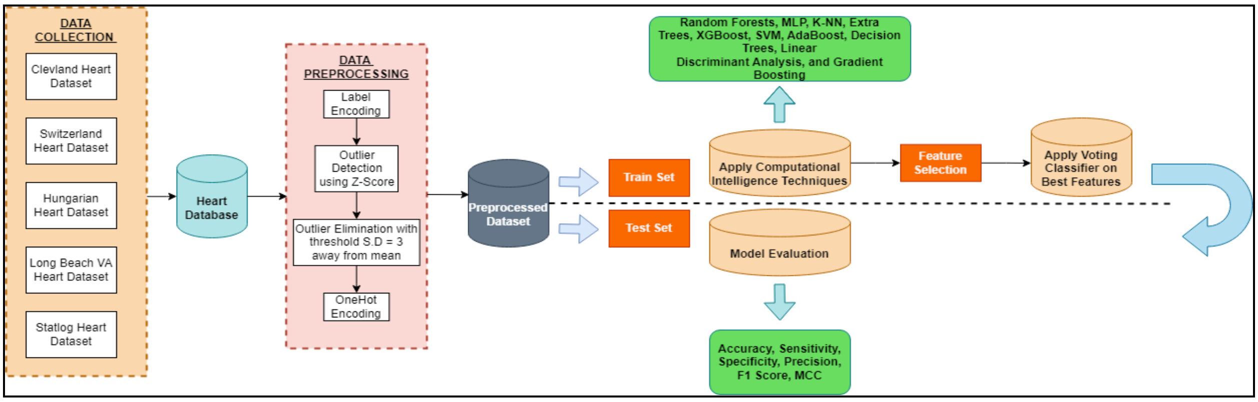

Figure 7: Full System for Coronary Heart Disease Diagnosis Prediction

V. PERFORMANCE MEASURE

Using the confusion matrix, we can visualize the performance of both the voting classifier models. In the confusion matrix, four classification performance indices are present. The definition of these is given below:

TP = True Positive (Correctly Identified)
TN = True Negative (Correctly Rejected)
FP = False Positive (Incorrectly Identified)
FN = False Negative (Incorrectly Rejected)

To evaluate the performance of the system, the following qualities are measured:

(1) Accuracy: Ratio of correctly classified samples to total samples. It is represented by:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (4)$$

(2) Sensitivity: Ratio of correctly classified positive samples to total positive instances, given by:

$$\text{Sensitivity} = TP / (TP + FN) \quad (5)$$

(3) Specificity: Ratio of correctly classified negative instances to total negative instances, given by:

$$\text{Specificity} = TN / (TN + FP) \quad (6)$$

(4) Precision: Precision is calculated as the number of correctly positive predictions divided by the total number of positively predict, given by:

$$\text{Precision} = TP / (TP + FP) \quad (7)$$

(5) F1 Score: The range of F1 scores is between 0 and 1. The best value is one, and the worst value is 0. F1 is represented as follows:

$$\text{F1 score} = 2TP / (2TP + FP + FN) \quad (8)$$

(6) Mathews Correlation Coefficient (MCC): For binary classification, the range of MCC is +1 to −1.

When the value is +1, the best performance is shown, and when the value is −1, the worst performance is shown. It is represented as:

$$\text{MCC} = (TP \times TN) - (FP \times FN) / \text{sqrt}((TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)) \quad (9)$$

VI. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. Experimental Setup

The experiment was conducted using the Python programming language on a single Virtual Machine configured through Microsoft Azure, with the following specifications: Intel(R) Xeon(R) Platinum 8272CL CPU @ 2.60 GHz and 32 GB RAM, running on the Windows 10 operating system. This information is supplied to aid understanding and replication of the findings.

B. Results Analysis

In this part, we give a performance assessment of several computational intelligence methodologies presented in Part III. We trained various machine learning models using the heart disease dataset, with a training-to-testing information ratio of 80:20. The accuracy ratings achieved by the different machine learning models are as follows: Random Forest (90.63%), Multi-Layer Perceptron (82.95%), K-Nearest Neighbors (80.68%), Extra Tree Classifier (89.77%), XGBoost (90.34%), Support Vector Classifier (82.67%), Stochastic Gradient Descent (73.86%), Adaboost (80.97%), and Classification and Regression Trees (82.95%). In addition to accuracy, we also examined the performance of the heart disease prediction model using numerous other measures, including precision, sensitivity, specificity, F1 score, and the Mathew correlation coefficient. The results of these measures are provided in Table 2.

We then integrate these different models into a hard voting classifier to increase the overall accuracy. The hard voting classifier aggregates the outcomes of each model and selects the class with the highest number of votes. After merging the separate models, we achieved an accuracy score of 90.34%, which is greater than the individual models. This implies that the mix of models has a beneficial influence on the overall accuracy of the classifier.

Table 2 : Evaluation Metrics of the Proposed Heart Disease Prediction System

Parameters	RF	MLP	KNN	SVC	GBM	CART	XGB	Voting Classifier
Sensitivity (%)	94.56	87.5	83.69	85.32	87.5	82.06	91.3	94.35
F1 Score	91.33	84.29	81.91	83.73	84.96	83.42	90.81	92.49
Accuracy (%)	90.62	82.95	80.68	82.67	83.8	82.95	90.34	91.91
Precision (%)	88.32	81.31	80.2	82.19	82.56	84.83	90.32	90.69
Specificity (%)	86.3	77.97	77.38	79.76	79.76	83.92	89.28	89.18
MCC	0.813	0.659	0.659	0.652	0.675	0.659	0.806	0.838

We sought to enhance the accuracy of the vote classifier through feature selection. Our feature selection strategy enhanced the accuracy of the voting classifier, achieving an accuracy score of 91.91%, which is higher than the original voting classifier's accuracy of 90.34%. This suggests that feature selection may be a valuable approach to enhance the classifier's accuracy.

Overall, our findings suggest that integrating separate models into an ensemble model and conducting feature selection can significantly enhance the accuracy of the heart disease prediction model. These results have significant implications for developing more accurate and reliable heart disease prediction models, which may aid in the diagnosis and treatment of heart disease.

To further show the efficacy of our final feature-selected voting classifier model, we have included a Receiver Operating Characteristic (ROC) curve plot in Figure 8. The ROC curve is a graphical depiction of the actual positive rate (sensitivity) versus the false positive rate (1-specificity) for various classification thresholds.

The area under the ROC curve (AUC) provides a single statistic to evaluate the model's performance, with values ranging from 0.5 to 1.0. An AUC value of 0.5 suggests random chance, whereas an AUC value of 1.0 shows flawless categorization.

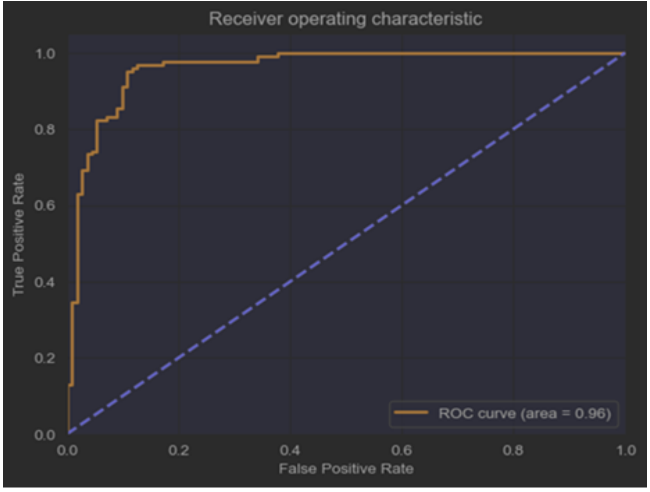

Figure 8: ROC Curve Plot for The Final Feature-Selected Voting Classifier Model.

As illustrated in Figure 8, our final feature-selected voting classifier model attained an AUC of 0.96, demonstrating outstanding performance in differentiating between positive and negative instances of heart disease. The steep curve and high AUC value indicate that the model has high sensitivity and specificity, suggesting it can reliably predict both genuine positives and true negatives.

Additionally, we included a Precision-Recall curve plot for the final feature-selected voting classifier model in Figure 9. The Precision-Recall curve is another essential assessment measure for binary classification problems, especially in situations where the classes are unbalanced. The curve is constructed by graphing the precision (positive predictive value) against the recall (actual positive rate) for various categorization thresholds. The area under the Precision-Recall curve (AUC-PR) provides a single statistic to evaluate the model's performance, with values ranging from 0 to 1.

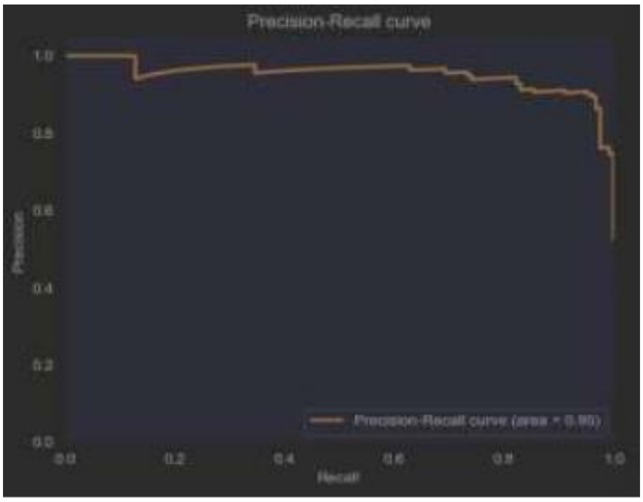

Figure 9: Precision-Recall Curve plot for the Final Feature-Selected Voting Classifier Model.

As illustrated in Figure 9, our final feature-selected voting classifier model achieved an AUC-PR of 0.95, demonstrating good accuracy in detecting positive instances of heart disease while minimising false optimistic predictions. The high accuracy and recall numbers reflect the model's ability to reliably identify positive cases, while the steep curve and high AUC-PR value imply a low false positive rate.

VII. CONCLUSION

In conclusion, our work demonstrates that combining various machine learning models into an ensemble model and utilising feature selection strategies can enhance the accuracy of heart disease prediction models. The individual models we examined, including Random Forest, Multi-Layer Perceptron, and XGBoost, demonstrated high accuracy rates; however, integrating them into a hard voting classifier significantly increased the overall accuracy. Moreover, adding feature selection methods to identify the most significant characteristics for prediction further enhanced the accuracy of the voting classifier.

Our results have significant implications for improving the detection and treatment of heart disease. With dependable prediction models, physicians and medical professionals may offer more effective treatment and preventative measures for individuals at risk of heart disease. The integration of multiple models and feature selection strategies can also be applied to other healthcare domains to enhance the accuracy of prediction models and support informed clinical decision-making.

While our study yields encouraging findings, several limitations remain that need to be addressed in future research. For instance, the dataset used in this research comprises only a limited number of characteristics, and it may be interesting to examine other features that could further increase the accuracy of the prediction models. Additionally, it would be helpful to study the generalizability of our technique across diverse datasets and demographics.

Overall, our work demonstrates the potential for combining machine learning models and employing feature selection strategies to enhance heart disease prediction accuracy. Our results provide a foundation for ongoing research and development of precise and dependable prediction models to improve clinical decision-making in the fight against heart disease.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to Dr. K. Ruba Soundar, Associate Professor in the CSE Department at Mepco Schlenk Engineering College, for his significant help, advice, and mentorship during our study endeavour. We are grateful for his ongoing encouragement, unwavering support, and valuable suggestions, which have significantly contributed to the successful completion of this task.

We would also like to acknowledge the contributions of the authors of the heart disease dataset used in this research, as well as the developers of the numerous machine learning tools and frameworks employed in this study. Their effort has made this research feasible and supplied us with the tools and resources required to perform this investigation.

Finally, we would like to thank our colleagues and friends who have given us vital comments, encouragement, and support throughout the study process. Their suggestions and thoughts have been incredibly helpful in constructing this study article.

DECLARATION STATEMENT

Funding	No, I did not receive it.
Conflicts of Interest	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval or consent to participate, as it presents evidence that is not subject to interpretation.
Availability of Data and Materials	Not relevant.
Authors Contributions	All authors have equal contributions to this article.

REFERENCES

1. World (2021). Cardiovascular diseases (CVDs). [online] Who.int. Available at: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) [Accessed 26 Apr. 2023].
2. CDC (2022). Heart Disease Facts. [online] Centres for Disease Control and Prevention. Available at: https://www.cdc.gov/heartdisease/facts.htm [Accessed 27 Apr. 2023].
3. Mayoclinic.org. (2022). Electrocardiogram (ECG or EKG) - Mayo Clinic. [online] Available at: https://www.mayoclinic.org/tests-procedures/ekg/about/pac-20384983 [Accessed 27 Apr. 2023].
4. Park, D.-J., Park, M.S., Lee, H.K., Young Ho Kim, Kim, Y. and Park, Y.-H. (2021). Development of a machine learning model for predicting disease diagnosis based on laboratory tests. Scientific Reports, [online] 11(1). https://doi.org/10.1038/s41598-021-87171-5
5. May, M.J. (2021). Eight ways machine learning is assisting medicine. Nature Medicine, [online] 27(1), pp.2–3. https://doi.org/10.1038/s41591-020-01197-2
6. M.A. Ganaie, Hu, M., Tanveer, M. and Ponnuthurai Nagaratnam Suganthan (2021). Ensemble deep learning: A review. Engineering Applications of Artificial Intelligence, [online] 115, pp.105151–105151. https://doi.org/10.1016/j.engappai.2022.105151
7. Uci.edu. (2023). UCI Machine Learning Repository: Heart Disease Data Set. [online] Available at: https://archive.ics.uci.edu/ml/datasets/Heart+Disease [Accessed 27 Apr. 2023].
8. Google Developers. (2022). Available at: https://developers.google.com/machine-learning/decision-forests/random-forests [Accessed 27 Apr. 2023].
9. MANU SIDDHARTHA (2020). Heart Disease Dataset (Comprehensive). [online] IEEE DataPort. Available at: https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive [Accessed 27 Apr. 2023].
10. Ibomoiye Domor Mienye, Sun, Y. and Wang, Z. (2020). An improved ensemble learning approach for the prediction of heart disease risk. Informatics in Medicine Unlocked, [online] 20, pp.100402–100402. https://doi.org/10.1016/j.imu.2020.100402
11. Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S. and Singh, P. (2021). Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning. Computational Intelligence and Neuroscience, 2021, pp.1–11. https://doi.org/10.1155/2021/8387680
12. Ahmad Bamanga Mahmud, Ahmadu Asabe Sandra, Musa Yusuf Malgwi and Sajoh, D.I. (2021). Ensemble model for Heart Disease Prediction. [online] Science Progress and Research. Available at: https://www.semanticscholar.org/paper/Ensemble-model-for-Heart-Disease-Prediction-Mahmud-Sandra/c941743125ea116d0862012b15330ce03ee8aa2d [Accessed 26 Apr. 2023].
13. Almulihi, A., Saleh, H., Ali Mohamed Hussien, Mostafa, S., Shaker El-Sappagh, Khaled Alnowaiser, Ali, A.A. and Moatamad Refaat Hassan (2022). Ensemble Learning Based on a Hybrid Deep Learning Model for Early Prediction of Heart Disease. Diagnostics, [online] 12(12), pp.3215–3215. https://doi.org/10.3390/diagnostics12123215
14. Naveed Anjum Itoo and Vijay Kumar Garg (2022). Heart Disease Prediction using a Stacked Ensemble of Supervised Machine Learning Classifiers. 2022 International Mobile and Embedded Technology Conference (MECON). [online]. https://doi.org/10.1109/MECON53876.2022.9751883
15. Kannan, S. (2022). An automated clinical decision support system for predicting cardiovascular disease using an ensemble learning approach. Concurrency and Computation: Practice and Experience, [online] 34(18). https://doi.org/10.1002/cpe.7007
16. Mahesh, T.R., Kumar, V., Kumar, V., Asghar, J., Oana Geman, G. Arulkumaran and Arun, N. (2022). AdaBoost Ensemble Methods Using K-Fold Cross-Validation for Survival with Early Detection of Heart Disease. Computational Intelligence and Neuroscience, [online] 2022, pp.1–11. https://doi.org/10.1155/2022/9005278
17. Ghulab Nabi Ahmad, None Shafiullah, Fatima, H., Abbas, M., Md. Obaidur Rahman, None Imdadullah and Alqahtani, M. (2022). Mixed Machine Learning Approach for Efficient Prediction of Human Heart Disease by Identifying the Numerical and Categorical Features. Applied sciences, [online] 12(15), pp.7449–7449. https://doi.org/10.3390/app12157449
18. Lubna Riyaz, Muheet Ahmed Butt and Zaman, M. (2022). Ensemble Learning for Predicting Coronary Heart Disease. 2022 2nd International Conference on Intelligent Technologies (CONIT). [online] https://doi.org/10.1109/CONIT55038.2022.9848292

AUTHORS PROFILE

Yogesh is a final-year student at Mepco Schlenk Engineering College with a keen interest in technology. Proficient in Java programming, he has completed multiple projects using Java Swing Classes. Currently, Yogesh is focused on Cloud Computing. He is an active member of tech and coding clubs, known for his dedication and innovation. His academic achievements and passion for technology promise a promising future in the field.

Paneer Thanu Swaroop is a dedicated student currently enrolled in the Computer Science department at Mepco Schlenk Engineering College, with a passion for technology. He has actively engaged in three impactful projects, showcasing his proficiency in Java programming. His project work reflects a keen interest in software development and a commitment to applying theoretical knowledge to real-world challenges.

Ruba Soundar K received the A.M.I.E. degree in Computer Science and Engineering from the Institution of Engineers (India) in 2000. He received the M.E. and Ph.D. degrees in Computer Science and Engineering from Anna University, Chennai, in 2004 and 2010, respectively. Currently, he is an Associate Professor (Sr. Grade) in the Computer Science and Engineering Department of Mepco Schlenk Engineering College, Sivakasi, Tamil Nadu. India. He has authored/coauthored over 120 research articles in various Journals and Conferences in the areas of Cloud Computing, Image Processing, Wireless and Wired Networking.
