Frontiers in Health Informatics
ISSN-Online: 2676-7104
www.healthinformaticsjournal.com
2024; Vol 13: Issue 5
Open Access

Forecasting Efficient Prediction of Diabetic Coronary Heart Disease Using Hybrid Ensemble Learning Method (HELM) and Feature Selection Algorithm

Mrs.S.Madhumalar[1] , Dr.S.Sivakumar[2]
[1]Assistant Professor, Cardamom Planters' Association College, Affiliated to Madurai Kamaraj University, Madurai
[2]Principal, Cardamom Planters' Association College, Affiliated to Madurai Kamaraj University, Madurai

Cite this paper as: Mrs.S.Madhumalar , Dr.S.Sivakumar(2024) Forecasting Efficient Prediction of Diabetic Coronary Heart Disease Using Hybrid Ensemble Learning Method (HELM) and Feature Selection Algorithm. *Frontiers in Health Informatics, 13* (5), 45-62

Abstract:
Artificial intelligence is a component of machine learning, which is utilized in data science to solve several issues. Predicting a result based on available data is a popular use of machine learning. Diabetic Coronary Heart Disease (DCHD) poses a significant healthcare challenge globally, demanding accurate predictive models for early detection and intervention. This study proposes a novel approach that integrates hybrid ensemble learning techniques with feature selection algorithms to enhance efficiency and accuracy of DCHD prediction. The hybrid ensemble learning method combines the multiple ensemble classifiers and compare with machine learning (ML)algorithms like decision trees (DT), support vector machine (SVM), and neural networks (NN), adaboost, XGboost to form a robust predictive model. Additionally, feature selection algorithms are employed to identify the most relevant predictors from a comprehensive set of clinical and demographic variables associated with DCHD. The proposed framework is validated using a large dataset comprising demographic information, medical history, and laboratory test results of diabetic patients. Performance evaluation metrics, including accuracy, sensitivity, specificity, and ROC curve, are utilized to assess the predictive performance of the model. Results demonstrate that the hybrid ensemble learning approach, coupled with feature selection algorithms, outperforms traditional single-model methods in terms of predictive accuracy and efficiency. The selected features provide valuable insights into the underlying risk factors and biomarkers associated with DCHD, facilitating early detection and personalized intervention strategies. Overall, this research contributes to the advancement of predictive modeling in healthcare by offering a sophisticated yet interpretable approach for forecasting DCHD risk in diabetic patients. The proposed framework has the potential to improve clinical decision-making and resource allocation, ultimately leading to better management and prevention of diabetic complications, including coronary heart disease.
Keywords: Ensemble learning, Accuracy, Machine learning, hybrid voting model, RUC- ROC curve, time complexity, computation time, feature selection.

1 INTRODUCTION

One of the best areas in which a large number of academics have demonstrated genuine interest is healthcare. Restructuring present healthcare procedures using various technology solutions is currently a major focus in order to give healthcare services that are both inexpensive and effective. Diabetic Coronary Heart Disease [1][2] (DCHD) refers to coronary heart disease (CHD) that occurs in individuals with diabetes mellitus. Plaque accumulates inside the coronary arteries, which provide oxygen-rich blood to the heart muscle, causing coronary heart disease (CHD), sometimes referred to as coronary artery disease (CAD). When someone with diabetes

develops CHD, it's termed Diabetic Coronary Heart Disease Diabetes, especially type 2 diabetes. The presence of diabetes can accelerate the formation of atherosclerosis (plaque build-up) within the coronary arteries, leading to CHD. The mechanisms linking diabetes and CHD are complex and multifactorial. Chronic high blood sugar levels in diabetes can lead to inflammation, endothelial dysfunction, and oxidative stress, all of which result in growth of atherosclerosis, subsequent CHD.

Individuals with diabetes have an increased risk of developing CHD compared to those without diabetes. Other risk factors for DCHD include hypertension, dyslipidaemia (abnormal levels of cholesterol and triglycerides), obesity, and a sedentary lifestyle. However, individuals with diabetes may also experience atypical symptoms or silent ischemia, where they do not experience typical chest pain despite having significant blockages in their coronary arteries. Diagnosis of DCHD involves a combination of clinical evaluation, imaging tests (such as coronary angiography), non-invasive tests (such as stress testing). Management strategies for DCHD include lifestyle changes, control of blood sugar levels, blood pressure management, lipid-lowering therapy, antiplatelet therapy, and in some cases, revascularization procedures like angioplasty or bypass surgery. DCHD poses a substantial healthcare burden due to its prevalence, complexity, and associated complications. It contributes to increased hospitalizations, and reduced quality of life for affected individuals. Given the high cardiovascular risk [2][3] associated with diabetes, prevention and management of DCHD are paramount. In essence, accurate prediction enables proactive decision-making, timely interventions, and resource optimization across various domains, ultimately leading to better outcomes, improved efficiency, and enhanced resilience in the face of uncertainties and challenges. Machine learning techniques [6][7] can potentially improve risk prediction by leveraging huge datasets and a variety of features. Various evaluation metrics is used to evaluate the performance of various machine learning classifiers and ensemble model in predicting the risk of DCHD. It also tries to improve the accuracy of predicting diabetic coronary HD using a strategy termed ensemble.

Therefore, the following offers a significant contribution to this developed heart disease diagnosis:

1. Ensemble learning method, combines multiple machine learning models to improve prediction accuracy. It often used in predictive modeling to aggregate the strengths of individual models.
2. A novel strategy that improves prediction accuracy by integrating feature selection algorithms with the hybrid ensemble learning method (HELM) is proposed for the early prediction of DCHD.
3. To reduce calculation time and improve prediction accuracy, the most valuable characteristics are extracted using the chi-square feature selection approach.
4. In order to forecast the patients' heart disease status based on their current state, the HELM was trained and learnt utilizing heart disease datasets such as Cleveland, Framingham, NHANES, MIMC, and Real time dataset.
5. Using assessment criteria including accuracy, specificity, precision, and F1 score, the suggested approach is assessed and contrasted with the findings of earlier research.
6. Furthermore, the statistical analysis that assessed the ensemble models' relevance in relation to other models was provided.

The remainder of this study is structured as follows: In Section 2, prior studies on the identification and prediction of DCHD using ensemble machine learning models are reviewed. Section 3 provides a detailed explanation of proposed HELM for DCHD prediction. The experimental results are presented in Section 4 along with a comparison with current approaches. Section 5 concludes by summarizing the results and offering suggestions for additional research.

2. Literature review

A significant global health concern, cardiovascular disease is thought to be the cause of 17.9 million deaths

annually. Machine learning and data analysis have become effective methods for identifying and averting cardiovascular disease. Researchers can find patterns and risk factors linked to CVD by analyzing complex medical records. They can also develop accurate and reliable predictive models by using machine learning models to find key features. Abhishek et al. (2023) demonstrated CatBoost classification algorithm for efficient prediction. To compare its efficiency imputation techniques, including Mean, Median, KNNI is used along with Hungarian dataset. Every day, massive amounts of patient data related to CVD are stored. This data can be preprocessed and used to train learning models, which will aid in the early prediction of illness. In order to help patients begin medication as soon as possible, Charkha et al. (2023) developed deep learning algorithms such as RNN (Recurrent Neural Network), whose type is LSTM (Long Short-Term Memory). The accuracy of the proposed work's prediction is tested instantaneously by comparing it with existing working algorithms.

Heart is the primary organ for supplying oxygen-rich blood to all body cells. Therefore, if there is a condition that leads to heart disease, it could be fatal. Diabetes and high blood pressure are the main causes of cardiovascular illnesses. A review of several machine learning techniques, including SVM, DT, ANN, Hidden Naïve Bayes, Naïve Bayes, Random Forest, K-Nearest Neighbors (KNN), and Particle Swarm Optimization (PSO), was conducted by Shree Raksha et al. in 2022. The article then discusses some of the issues surrounding cardiovascular illness, how characteristics affect the result, and how machine learning are used to forecast the condition. By using the output of a given dataset, many lives can be spared. An ensemble framework for the early prediction of DCHD was presented by Chowdary et al. (2023). The approach makes use of data pre-processing, outlier identification, and predictive machine learning classifiers such XGBoost (XGB), Naïve Bayes (NB), SVM, k-Nearest Neighbors (k-NN). The grid search method is used to fine-tune base classifiers to improve prediction.

Talapaneni et al. (2024) propose a proficient model for prediction of CVD using a Voting Ensemble technique. It combines Logistic Regression, DT, and Random Forest classifiers to create a robust predictive model. Real-world heart disease dataset is used measure its accuracy, performance. Voting Ensemble model outperformed individual classifiers, achieving an exceptional 98% accuracy on the test dataset. Chakraborty et al (2024) demonstrated the post-processing ensemble Machine Learning (ML) for classification of CVD. Data is collected from publicly available Kaggle dataset, then missing values are handled and data unbalancing techniques are used for the data pre-processing. After that, Principal Component Analysis (PCA) is used for selection of relevant features. Finally, post-processing ensemble ML approach is used for the classification of CVD. It is estimated by various performance metrices and it achieves high accuracy when compared to the previous approaches such as SVM and Gradient Boosting (GB).

Vinora et al. (2023) demonstrated ensemble stacked model which combine SVM, DT models to predict the disease accurately. The dataset form Kaggle is used for prediction. Review existing prediction models and their limitations. Reshan et al. (2023) proposed Hybrid Deep Neural Networks (HDNNs) an innovative approach, which combine CNN, LSTM, and dense layers for heart disease prediction and highlight their performance. It looks at three deep learning models: Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Long Short- Term Memory (LSTM). Emphasize the superior accuracy of HDNNs compared to traditional methods, as demonstrated on public datasets. While the HDNN system shows improved accuracy, there is an opportunity to explore its scalability and real-world applicability. Two publicly available heart disease datasets, including Cleveland dataset is used to test proposed models. To measure the system’s performance against conventional systems various metrics is used.

Mondal et al. (2024) demonstrates a dual-stage stacked machine learning model for predicting heart diseases, leveraging a dataset of 1190 patients with eleven significant characteristics. Five machine learning

classifiers were used, with hyperparameter tuning via RandomizedSearchCV and Grid- SearchCV. The best-performing models were refined using a stacking ensemble technique. The model demonstrated high stability with a comparable dataset. Dual-stage stacked model shows potential for accurate and prompt diagnosis of heart diseases, aiding in reducing global fatalities. Baghdadi et al. (2023) proposed Catboost model for automatically electing key features in early-stage heart disease detection, which could optimize early prediction and intervention. It involves in comprehensive data analysis, including feature selection and model evaluation.

Subramani et al. (2023) discusses cardiovascular disease (CVD) prediction using ML and deep learning, emphasizing the need for AI technologies to predict outcomes in CVD patients. It uses a stacking model with a base and meta-learner layers. For feature selection, it uses the SHAP method and classifiers such as Random Forest (RF), Logistic Regression (LR), Multilayer Perceptron (MLP), Extra Trees (ET), CatBoost, and Gradient Boosting Decision Tree. It highlights potential of AI-based technologies, particularly Internet of Things (IoT), in enhancing the prediction of CVD outcomes. By accounting for complex data interactions, machine learning models can significantly improve over traditional statistical models. Incorporated method achieved nearly 96% accuracy in predicting CVD, outperforming existing methods. According to the study, deep learning could benefit from more medical institution data for developing artificial neural network structures. Sarra et al. (2022) discusses machine learning (ML) advancements for CVD prediction. It highlights using SVM algorithm and a chi-squared ($\chi 2$) statistical method for feature selection to enhance prediction accuracy. It references studies on the global impact of heart disease, emphasizing the need for accurate prediction and early diagnosis to reduce mortality rates. It underscores the significance of selecting relevant features in ML models to avoid overfitting and improve performance. The SVM model with $\chi 2$ feature selection is evaluated against traditional models using different metrics demonstrating improved prediction capabilities.

Rahim et al. (2021) presents the Machine Learning based Cardiovascular Disease Diagnosis framework, which uses highly precise machine learning techniques to predict cardiovascular diseases. It addresses missing values with mean replacement and data imbalances with Synthetic Minority Over-sampling Technique, ensuring data reliability for accurate predictions. Then, for feature selection, the Feature Importance approach is applied. Lastly, a combination of KNN and logistic regression classifiers is suggested to provide a more accurate prediction. Framingham, Heart Disease, and Cleveland are used to validate the framework. Ultimately, the comparison analysis demonstrates that MaLCaDD predictions outperform current methods in terms of accuracy (but using a smaller feature set). MaLCaDD is therefore extremely dependable and suitable for use in real-world settings for the early detection of cardiovascular illnesses. Elsedimy et al. (2023) describes a new method to identify heart disease. It uses a quantum-behaved particle swarm optimization (QPSO) and support vector machine (SVM) classification model aims to enhance the accuracy of cardiovascular disease predictions. The use of QPSO for feature selection in CHD prediction is innovative, as it enhances the SVM's ability to handle high-dimensional data. The model's self-adaptive threshold method for parameter tuning is unique, helping to avoid local minima and ensuring a balance between exploration and exploitation. When applied to Cleveland dataset, QPSO-SVM model achieved high prediction accuracy and outperformed other models.

3. Materials and Methods

The sections that follow provide an explanation of how the suggested HEML model for heart disease detection operates. The steps involved in developing an ensemble model are explained in detail in this section, starting with feature scaling and data collecting and ending with the selection and assessment of suitable machine-learning models. Figure 1 shows the detailed architecture of hybrid ensemble learning method.

3.1 Description of the datasets:

3.1.1. Cleveland and Framingham dataset

The University of California, Irvine (UCI) repository provided the Cleveland dataset [23], while the Kaggle website provided the Framingham dataset [24]. In contrast to the Framingham, which has 4238 occurrences and 16 attributes, the Cleveland has 303 instances and 14 attributes. Health and demographic data, including age, sex, blood pressure, cholesterol, alcohol consumption, diabetes, and so on, are included in both datasets. Based on the input qualities, the output feature—basically a label attribute—is presented in the final column. For instance, value of 0 indicates that the patient is HD negative, and a value of 1 indicates that the patient is HD positive.

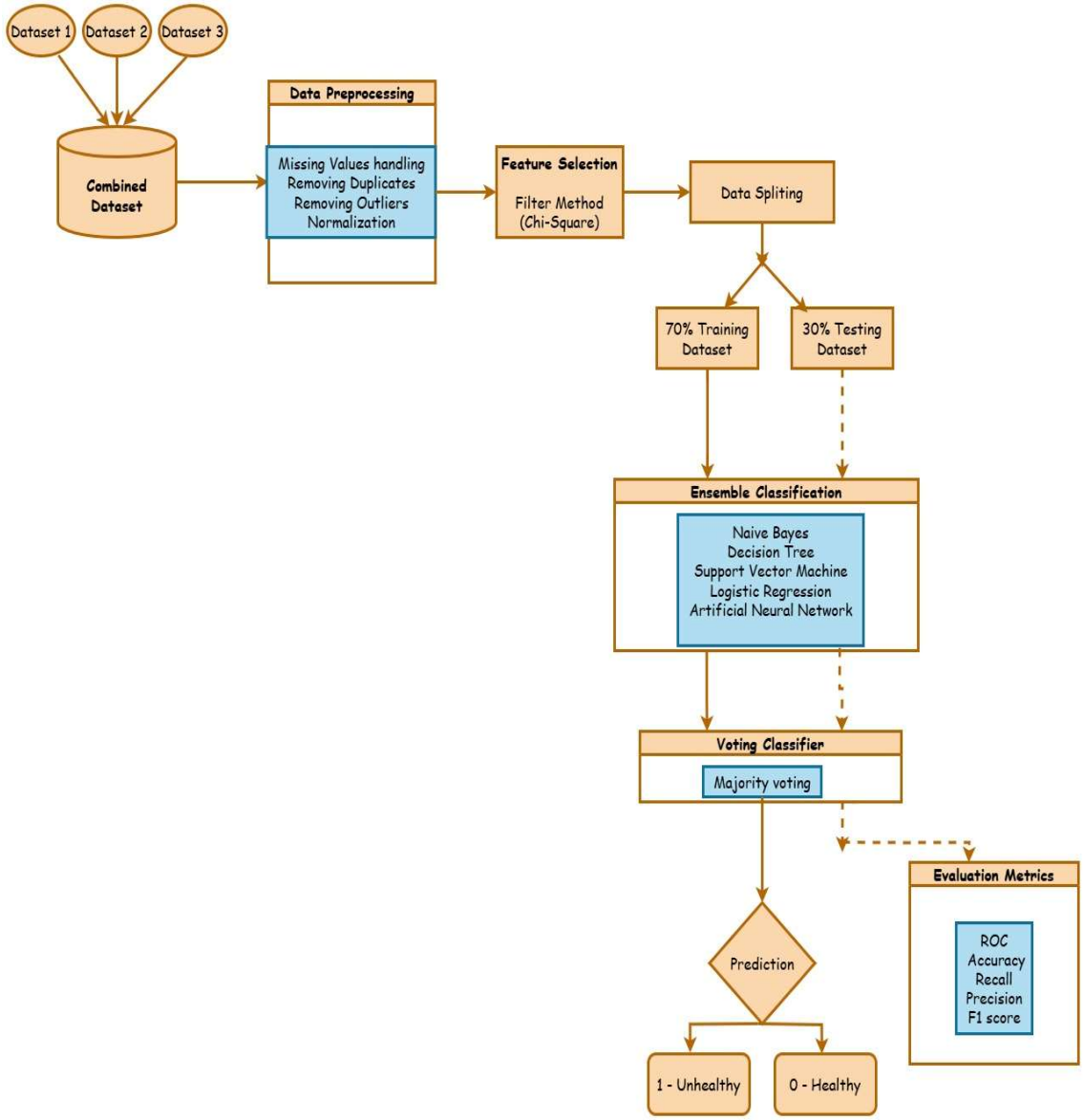

Figure 1. Proposed method Architecture

3.1.2. NHANES dataset

One of the main resources for tracking hypertension in the American population is National Health and Nutrition Examination Survey (NHANES) [25], which is carried out by National Center for Health Statistics (NCHS), and provides a wealth of features and data pertaining to cardiovascular diseases. The nationally

representative survey selects a sample of citizens from each of the 50 states as well as Washington, DC, using a sophisticated, multistage probability methodology. Every two years, participants are chosen at random to participate in laboratory testing, physical examinations, and home interviews. 33 482 people, ranging in age from 20 to 80, finished the tests and questionnaires. The 653 women who indicated they were pregnant at the time of the survey were not included. In the final analysis total of 10 212 participants (5301 males and 4911 women) with finished the interview and examination data.

3.1.3. MIMIC dataset

Medical Information Mart for Intensive Care (MIMIC-IV) [26] is a sizable deidentified dataset of patients admitted to the Boston (BIDMC) Beth Israel Deaconess Medical Center's emergency room or intensive care unit. It provides data above 65,000 patients hospitalized to an ICU and above 200,000 patients admitted to the emergency department. All medical record numbers for patients hospitalized to an ICU or emergency room between 2008 and 2022 were compiled into a master patient list. It includes comprehensive data on the vital signs, medications, test results, clinical comments, and patient demographics, spanning a range of hospital intensive care unit (ICU) stays. By providing more recent data, a larger patient group, and more record granularity than the prior MIMIC-III dataset, this version is superior. MIMIC-IV is divided into modules such as "hosp," which provides data on hospital admissions and treatments outside of intensive care units, and "core," which includes patient demographics.

3.1.4. Real time dataset

Diabetes and coronary heart disease patients' health information, such as blood pressure, cholesterol, heart rate, glucose levels, and lifestyle choices like exercise and food, will be continuously collected and updated in a real-time dataset. This dataset has the potential to combine information from mobile health apps, wearable technology, and electronic health records (EHR) to provide real-time insights into patient status. Predictive analytics for early diagnosis, risk evaluation, and customized treatment regimens would be supported. Such a dataset could help with early warning sign detection, better outcomes, and timely decision-making by healthcare providers for managing various chronic diseases with machine learning models.

3.2 Data Pre-processing

Preprocessing techniques are used to improve the performance of the model. These techniques include handling missing variables and eliminating duplicates and outliers. Extensive verifications and adjustments were carried out to guarantee the excellence and appropriateness of the amalgamated dataset. The dataset was first carefully checked to make sure there were no cases of missing data, and the findings indicated that there were none. This proved the integrity and trustworthiness of the dataset. Second, a careful review of duplicate values was done to guarantee data consistency because dataset was generated by combining several datasets. The lack of duplicates was verified by this study, supporting the dataset's accuracy. An outlier analysis was also performed in order to identify any extreme values that can potentially affect the data. Interestingly, there were no outliers found, demonstrating the dataset's resilience. Normalization aims to standardize pixel intensity values, ensuring consistent interpretation across scans. Min-Max normalization scales features to a range between 0 and 1, maintaining the relative differences between data points shown in Equation 1.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad (1)$$

WhereX_{min}, X_{max}, and X_{norm}represent the minimum, maximum, and normalized feature values, respectively, in the dataset, and X represents the original feature value.

3.3 Dataset Splitting

The method of data separation played a key part in this investigation. The dataset was divided into 80% training

set and 20% testing set in order to guarantee the creation of a reliable and accurate heart disease prediction model. During training, each base classifier undergoes adjustments based on training errors utilizing 10-fold cross-validation with three repetitions. This rigorous validation process ensures robust model performance. With an appropriate measure guiding adjustments, the classifiers are refined to effectively capture the dataset's nuances and deliver reliable predictions. The research methodology is outlined in Figure 4, providing a comprehensive analysis of the dataset and algorithm performance.

3.4 Classification Models

Using already-existing data, classification is a supervised learning process that forecasts the result. This research suggests a method for improving classification accuracy by employing an ensemble of classifiers and a classification algorithm to predict diabetic coronary heart disease. Each classifier is trained using the training dataset once the dataset has been split into a test set and a training set. Using the test dataset, the classifiers' efficacy is evaluated.

3.4.1 Naive Bayes (NB)

Based on the assumption of conditional independence between features, the Naive Bayes classifier is the most straightforward, efficient, and probabilistic algorithm. It applies the Bayes theorem. Conditional independence denotes the absence of reliance between one feature value's changes and the other feature values within a class. Based on the likelihood that a specific object would be encountered by the classifier, predictions are generated. This makes it possible for NB algorithm to separately compute feature distributions P(A | B). The naïve Bayes classifier is helpful for classifying high dimensional datasets because it prevents issues brought on by high dimensionality by decoupling the feature distributions shown in figure 2.

$$P(A|B) = \frac{P(A|B)*P(B)}{P(A)} \quad (2)$$

Where $P(A|B)$ is the Posterior Probability of class A given features B, $P(B|A)$ is likelihood of feature set B given class A, P(A) is the Prior probability of class A, P(B): Marginal probability of feature set B.

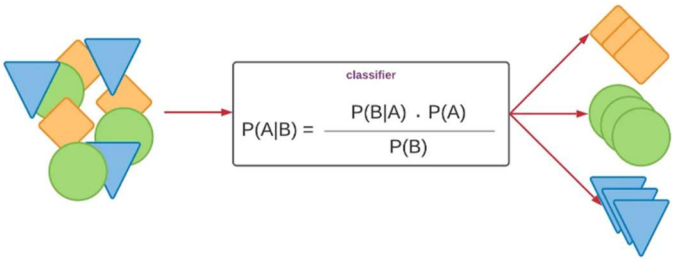

Figure 2. Naive Bayes

3.4.2 Decision Tree

An essential algorithm for supervised techniques is the decision tree shown in figure 3. Researchers use different tree-based classification methods, such Random Forests or Gradient Boosting, for different kinds of jobs. The training data is initially present in the root node, which uses the Gini Index to determine which qualities are the best. The two main stages of building any decision tree algorithm are (i) tree growth, which divides the training set repeatedly according to local optimal criteria until most of the data in the segment have the identical class label and (ii) tree pruning, which minimizes the tree's size to make it more comprehensible. The algorithm determines which attribute offers the greatest reduction in uncertainty at each node of the trees by assessing the entropy and Gini impurity of each attribute. The following formulas were used to determine the entropy and Gini impurity:

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

$entropy = -\sum_{i=1} P_i log_2(P_i)$ (4)

where P_i represents the probability of class i, i is the total number of classes.

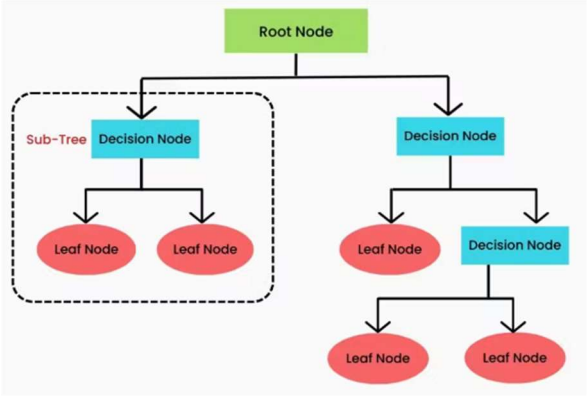

Figure 3. Decision Tree

3.4.3 Support Vector Machine (SVM)

One of the most well-known and useful methods for addressing problems with data classification, learning, and prediction is support vector machines (SVM) shown in figure 4. Support vectors are the data points that are closest to the decision surface. In infinite dimensional space, it classifies data vectors using a hyperplane. The maximal margin classifier, which helps identify the most fundamental classification problem of linearly separable training data with binary classification, is the most basic type of SVM. In real-world complexity, the hyperplane with the largest margin is identified by the maximal margin classifier. The solution to the SVM optimization problem is to minimize:

$\frac{1}{2} \| w \|^2$ (5)

$y_i(w \cdot x_i + b) \geq 1$ (6)

Where w is the weight vector, b is the bias, x_i represents the input features, and y_i are the class labels.

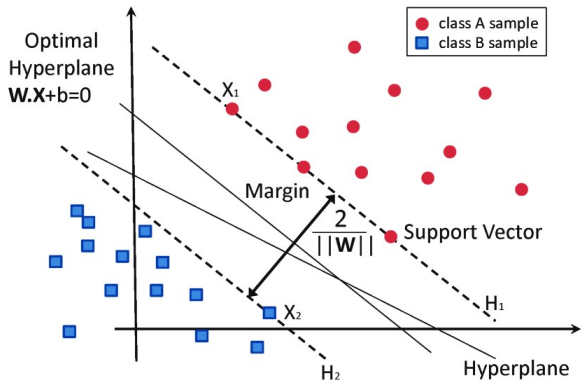

Figure 4. Support Vector Machine

3.4.4 Logistic Regression (LR)

A popular machine learning model for regression tasks is called LR shown in figure 5. On the other hand, this approach is also frequently applied to binary classification problems, which predict the likelihood of a categorical target variable. It can be used for multiclass classification using multinomial logistic regression, although it is most effective for binary classification issues. In order to categorize the data, the LR model fits the parameters using the maximum likelihood function with gradient descent, assuming that the data have a Bernoulli distribution. It makes use of a non-linear function to determine the type of fresh input, such as the

logistic or sigmoid functions. The generic version of the logistic regression model is represented by the following statement.

$$P(Y-1/X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots\ldots+\beta_n X_n)}} \quad (7)$$

where Y - binary outcome variable; β_0 - intercept coefficient; $\beta_1, \beta_2, \ldots, \beta_n$ are coefficients associated with features $X_1, X_2, \ldots, X_n$.

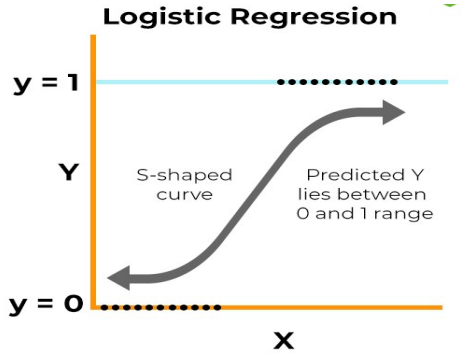

Figure 5. Logistic Regression

3.4.5 Artificial Neural Networks

Artificial Neural Networks (ANNs) are computing systems inspired by the human brain's network of neurons shown in figure 6. They consist of interconnected nodes or "neurons," arranged in layers: input, hidden, and output. The formula for an ANN's output is:

$$y = f(\sum_{i=1}^{n} w_i x_i + b) \quad (8)$$

Where y is the output, x_i are the input features, w_i are the weights, b is the bias term, and f is the activation function (e.g., sigmoid, ReLU). This formula helps the network learn and model complex patterns in data.

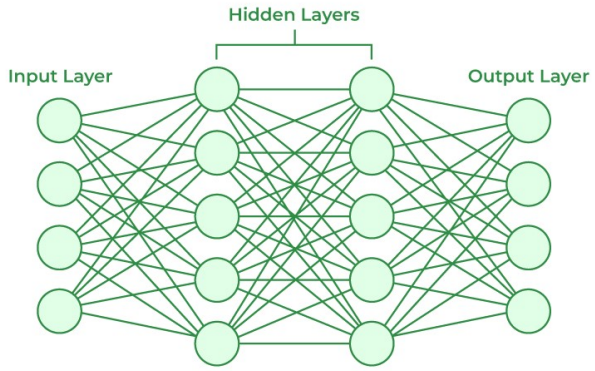

Figure 6. Artificial Neural Networks

3.6 Ensemble Techniques

Ensemble methods are powerful machine learning method that combine predictions of different models to improve accuracy and robustness. Rather than depending just on one model, ensembles leverage strengths of various algorithms or different versions of the same algorithm. Common ensemble techniques include bagging, boosting, stacking, and voting. Bagging reduces variance by averaging predictions, while boosting focuses on correcting errors by sequentially training models. Stacking combines predictions through a meta-learner, and voting aggregates predictions using majority rules. These methods help reduce overfitting, enhance predictive performance, and provide more stable models in complex machine learning tasks.

3.6.1 Bagging

The machine learning ensemble meta-algorithm known as bagging or bootstrap aggregating. It was created to increase stability, precision of ML algorithms used in statistical regression and classification. Using replacement, bagging chooses a random subset of patterns from the training set. To provide the final outcome, the performance of each classifier is added up. This method, which has low bias and large variance, is applied to weak learners. This approach consists of three steps: bootstrapping, parallel training, and aggregation. At first, different subsets of the data are created by selecting and replacing data points at random. These data sets are referred to as bootstrap replicates. Subsequently, these data subsets undergo individual and concurrent training. Lastly, averaging or majority voting is used to combine the output from each classifier. Implementing bagging is simple and reduces variance.

Algorithm 1 Bagging Algorithm

Input: training data $S = (x_1, y_1), \ldots, (x_2, y_2), \ldots, (x_n, y_n)$
Base ML algorithm L
The number of base learners T.
Procedure:
for $t = 1, \ldots, T$:
1) Generate a bootstrap sample S_j from the input data S
2) Fit a base learner h_j using S_j, i.e. $h_j = L(S_j)$
end for
Output: Combine the outputs of the base learners, $H(x) = mode(h_1(x), \ldots, h_T(x))$

3.6.2 Boosting

Boosting is an ensemble machine learning technique that builds a strong learner by combining several weak learners. It operates by training weak models one after the other, with each new model concentrating on the errors caused by the preceding ones. Boosting techniques, in example, employ input data to train a weak learner, calculate learner's predictions, choose training samples that have been incorrectly classified, and train a subsequent weak learner using an altered training set that includes the incorrectly classified instances from the previous training round. Until a predetermined number of basis learners is obtained, the iterative learning process is continued, and the base learners are then weighted collectively. Reducing bias is the main goal of boosting rather than variance. The gradient boosting machines (GBM) are utilized in this work to detect heart problems. In GBM, all estimators are added gradually by adjusting the weights, and weak learners are trained one after the other. The gradient boosting algorithm seeks to minimize the discrepancy between the projected and actual values by anticipating the residual errors of earlier estimators.

Algorithm 3 Boosting Algorithm (Gradient Boosting)

Input:
a training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$
a differentiable loss function $L(y, \hat{y})$
a base learner algorithm
number of boosting iterations M

1: Initialize the model with a constant value:

$$F_0(\mathbf{x}) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$$

2: **for** $m = 1$ to M **do**
3: Compute the pseudo-residuals:

$$\tilde{y}_{im} = -\frac{\partial L(y_i, F_{m-1}(\mathbf{x}_i))}{\partial F_{m-1}(\mathbf{x}_i)}, \quad i = 1, \ldots, n$$

4: Fit a base learner $h_m(\mathbf{x})$ to the training set $\{(\mathbf{x}_i, \tilde{y}_{im})\}_{i=1}^{n}$
5: Update the model:

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + h_m(\mathbf{x})$$

6: Output the final ensemble $F_M(\mathbf{x})$

3.6.3 Stacking

Stacking is an ensemble technique where a meta classifier is used to integrate numerous classification models. One layer is layered upon another, with each model feeding its predictions to the one above it, and the top-level model making decisions depending on the models below it. The original dataset provides input features to the models at the bottom layer. Using the output from the bottom layer, the top layer model predicts. The process typically involves two stages. In the first stage, several diverse models are trained on the same dataset. These models can be of different types (e.g., decision trees, support vector machines, neural networks), which helps capture various patterns in the data. In the second stage, the predictions of these base learners are fed into a meta-model, which learns how to best combine them to produce the final prediction. The key advantage of stacking is that the meta-learner can improve upon the weaknesses of the base learners, often resulting in better accuracy.

Algorithm 3 Stacking Algorithm

Input: training data $S = (x_1, y_1), \ldots, (x_2, y_2), \ldots, (x_m, y_m)$
The base learning algorithms T
Procedure:
Step 1: Train base learning models
for $t = 1, \ldots, T$:
Fit a base learner h_t using S
end for
Step 2: Obtain a new dataset from S
for $t = 1, \ldots, T$:
Obtain a new dataset containing $\{\hat{x}_i, y_i\}$, where $\hat{x}_i = \{h_1(x_i), h_2(x_2), \ldots, h_T(x_i)\}$
end for
Step 3: Train the meta-learner $\hat{h}$ using the new dataset
return $H(x) = \hat{h}(h_1(x), h_2(2), \ldots, h_T(x))$
Output: A stacked ensemble classifier H.

3.6.4 Majority voting

Majority voting is an ensemble learning technique used in classification and regression tasks where multiple models (or classifiers) predict the class label of an input. The class that obtains the majority of votes from different models makes the final prediction. In the meanwhile, regression tasks acquire the majority vote by averaging the predictions made by each base learner. This creates a single ensemble model that is trained by separate classifiers to forecast the output for each output class according to the majority of votes cast collectively. It is a simple yet powerful method to aggregate the predictions from different models, often leading to better performance than individual models. The final class label d_j is defined as

$$d_j = mode\{c_1, c_2, \ldots., c_n\} \quad (10)$$

Where $\{c_1, c_2, \ldots., c_n\}$ represents the individual classifiers that participate in the voting.

Algorithm 4 Majority Voting algorithm

Input:

D: Training dataset with labels representing C Class

L: Learning algorithm

W: Labels of the training dataset

N: Number of L used

Do n=1 to N

1. Call L with D_n and receive the classifier L_n.
2. Compare Wn with Cn generated from Ln, update vote.
3. Aggregate vote to the ensemble

End

3.7 Enhanced Ensemble Model with Feature Selection (FS)

selecting characteristics from a large pool of available attributes or features is the process of feature selection, which aims to increase accuracy while decreasing computing complexity and delay. To enhance model performance, choosing the appropriate features for training and testing data is crucial. A higher score indicates that the feature is more significant or appropriate. Prior to using the ensemble model, the key features are chosen using the Chi-square statistical algorithm model.

Utilizing a chi-square statistical approach, the most significant features identified and bias from the training set is removed. It assesses each feature's degree of independence from the target variable and assigns a score based on how strongly they are associated. It establishes the level of correlation between the predicted class and the input information. It is used to determine which features depend on the projected class for each non-negative feature (b_i). A rising chi-square value suggests that the feature depends heavily on the anticipated class shown in table 1. The chi-square test is used to rank the features of a binary classification issue in the following manner: A positive and negative set of class outputs and a total of (t) instances are assumed.

Table 1. Chi-square test score calculation

	Class (+ve)	**Class (-ve)**	**Total**
When feature b_i is present	e	k	m = e + k
When feature b_i is absent	f	l	t – m = f + l
Total	g = e + f	t – g = k + l	t

Where (t - g) represents total number of non-positive examples; (m) represents total number of instances where x_i is present, (t - m) is total number of instances where (b_i) is absent. In this (g) is sum of positive instances. The observed or actual count (O) is compared to expected or predicted count (E) using chi-square test. When two features are independent, actual, predicted counts are quite near. Let the values that were measured be represented by e, f, k, and l, and the projected values by E_e, E_f, E_k, and E_l. The expected value (E_e) can then be found using Equation (11) assuming there is no relationship between the two events. E_l, E_f, and E_k are also calculated. Finally, Equation (13) is used to calculate the chi-square score. Equation (12) provides the universal chi-square test form.

$$E_e = (e+f)X \frac{(e+f)}{t} \quad (11)$$

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (12)$$

$$\chi^2 = \frac{(e-E_e)^2}{E_e} + \frac{(f-E_f)^2}{E_f} + \frac{(k-E_k)^2}{E_k} + \frac{(l-E_l)^2}{E_l} \quad (13)$$

The 13 features were ranked from highest to lowest importance using this FS on the Cleveland dataset, as illustrated in Figure 7. The six key characteristics for diagnosing HD were identified as follows: thalach, oldpeak, ca, cp, exang, and chol. The number of features used for diagnosis is cut in half by half when the chi-square approach is applied, from thirteen to six.

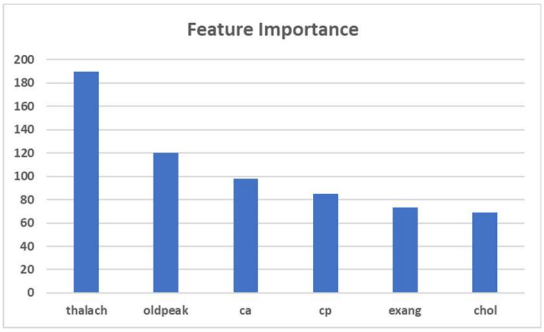

Figure 7. Features' importance according to chi-square scores

4. Experiments and results

4.1 Comparison with existing work

The efficiency of the suggested models was evaluated based on five criteria: Receiver Operating Characteristic (ROC), accuracy, recall, precision, and F1 score. true negative (TN) denotes accuracy of the algorithm's predictions for individuals without HD; true positive (TP) denotes accuracy of the algorithm's predictions for patients with HD; false positive (FP) denotes error-prone classification of patients without HD as having HD; and false negative (FN) denotes error-prone classification of patients with HD as healthy. The proposed method is compared against machine learning algorithms like decision trees, support vector machines, neural networks, adaboost, and XGboost. These metrics serve to identify the most effective system in predicting

DCHD.

$$\text{Accuracy} = \frac{\text{TP+T}}{\text{TP+TN+FP+F}} \quad (14)$$

$$\text{Sensitivity / Recall} = \frac{TP}{TP+} \quad (15)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (16)$$

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (17)$$

4.2 performance enhancement using feature selection

In the assessment, two methods were investigated. The first method involves training and assessing the hybrid ensemble models directly, without the use of the chi-square FS methodology, after standardizing the Cleveland dataset, which had 13 input characteristics. The majority voting ensemble classifier was then created by combining the prediction outputs of the base models. The ensemble classifier creates its own prediction after ensuring that the basis models' predictions are error-free. As a result, the HD diagnosis system was able to reduce the amount of error generated by each base classifier individually and increase its total accuracy. Figure 8 shows metrics evaluation before applying feature selection.

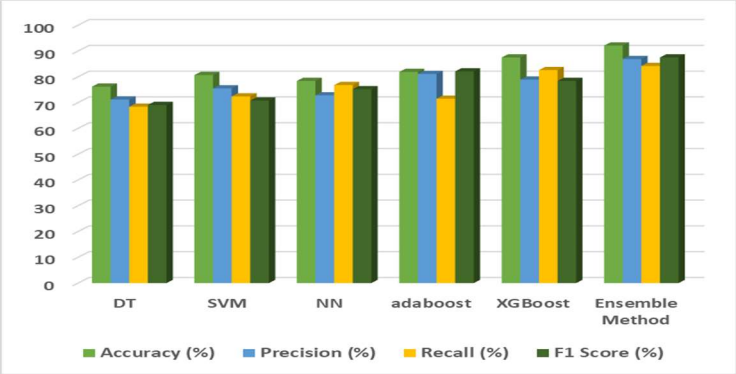

Figure 8. Metrics Before feature selection

In the second way, the chi-square FS algorithm was used to choose the five essential features following the dataset's normalization via the feature scaling method. The reduced-feature dataset was then used to train and assess hybrid ensemble models. Final prediction was generated by feeding the majority voting ensemble classifier with the prediction outputs from the basis models. Figure 9 shows metrics evaluation after applying feature selection. Applying feature selection improved the accuracy, precision, recall, and F1 score by 3.9%, 3.1%, 3.8%, and 2.5%, respectively.

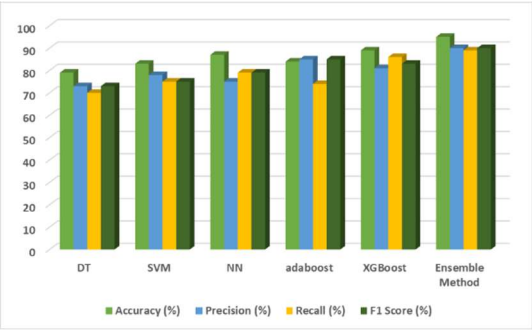

Figure 9. Metrics After FS

Also, area under the curve (AUC) and receiver operating characteristic (ROC) charts were utilized to

carry out a more thorough assessment of the diagnostic model's effectiveness. It evaluates the model's capacity to distinguish between two classes—0 denotes no HD and 1 denotes HD. Figure 10 shows ROC-AUC chart of the ensemble method without feature selection.

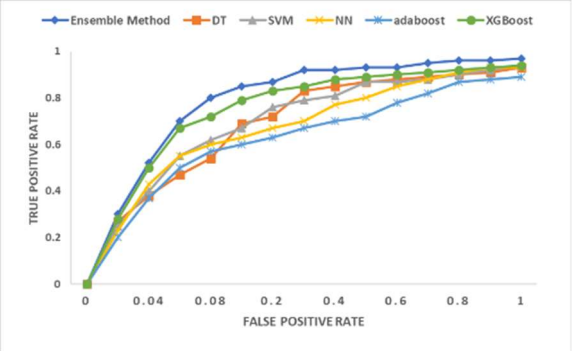

Figure 10. ROC- AUC chart before FS

A model that displays a superior ability to distinguish between the classes is the one with the curve in the top left corner of the graph, which has a lower false positive rate and a higher true positive rate. The model is producing accurate predictions ranges from 0 to 1. The model has the strongest separability when its AUC is close to 1, and the worst disassociation when it is close to 0. The ensemble method's ROC-AUC curve with feature selection is displayed in Figure 11.

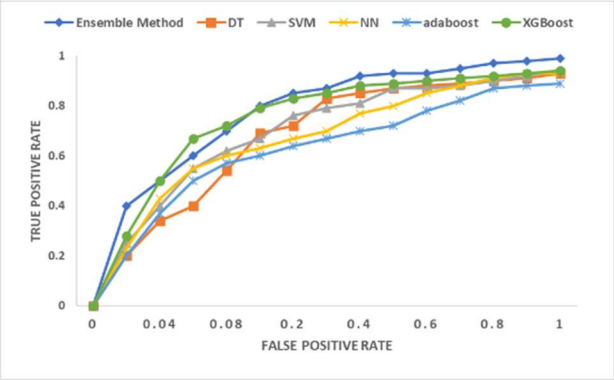

Figure 11 ROC- AUC chart after feature selection

Another statistic, confusion matrix is used to evaluate the efficacy of the prediction model even further. The ratio of right to wrong guesses is summarized in confusion matrix.

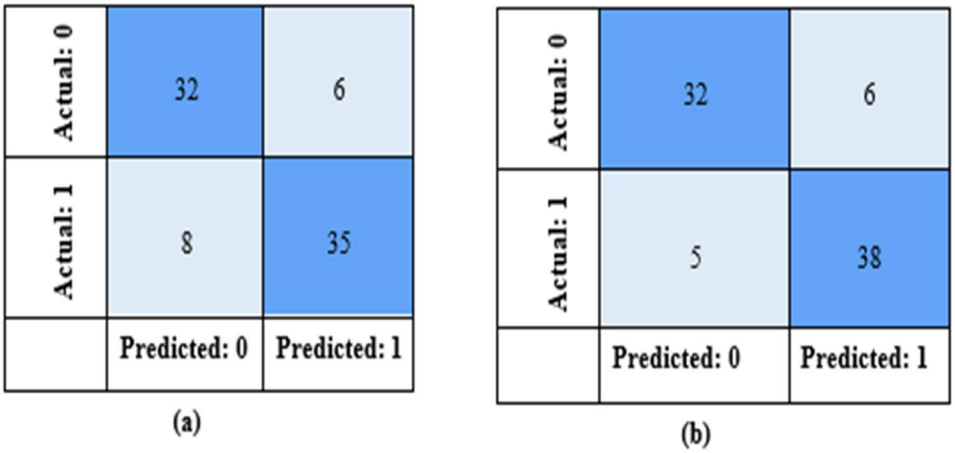

Figure 12. Confusion matrix (**a**) Before applying FS (b) after applying FS

Additionally, as shown in Figures 12(a), (b), the majority voting hybrid ensemble classifier's confusion matrix was calculated both before and after the FS procedure. The results demonstrate the value of feature selection in improving predictive models, with the recommended ensemble classifier exhibiting the best overall performance in early heart disease prediction.

5.Conclusion and future work

In many developing and most developed nations, DCHD is a leading cause of death. The substantial handicap caused by the clinical consequences of DCHD is a major contributor to the rising expenditures of healthcare. The purpose of this work is to present an efficient detection system for the detection of heart disease, based on the chi-square feature selection method and an ensemble technique. Various foundation machine learning models, such as LR, ANN, naïve Bayes (NB), DT, and support vector machine (SVM), are incorporated into the ensemble model. The ensemble models were trained and tested using the dataset on heart disease. Analysis was done on performance metrics such receiver operating curve, F1-score, recall, precision, and accuracy of classification. Six relevant features were found using the chi-square FS method, which improved majority voting ensemble model's performance and accuracy to an astounding 92.11%, with sensitivity of 96.61%, specificity of 90.48%, precision of 95.00%, and an F1 score of 92.68%. The majority voting ensemble model considerably outperforms the state-of-the-art methods for DCHD detection, according to experimental results.

In future research, explore the integration of deep learning techniques with the proposed Hybrid Ensemble Learning Method (HELM) to enhance the prediction accuracy of diabetic coronary heart disease. Additionally, more advanced feature selection algorithms, such as genetic algorithms and deep feature selection should be implemented, to improve model performance and reduce computational complexity. Expanding the dataset and applying cross-validation methods across diverse populations will also be prioritized for better generalization and clinical applicability.

References

1. Deshmukh, V.M. Heart disease prediction using ensemble methods. Int. J. Recent Technol. Eng. **2019**, 8, 8521–8526.
2. Sharma, R.; Singh, S.N. Towards Accurate Heart Disease Prediction System: An Enhanced Machine Learning Approach. Int. J. Perform. Eng. **2022**, 18, 136–148.

3. Aliyar Vellameeran, F.; Brindha, T. A new variant of deep belief network assisted with optimal feature selection for heart disease diagnosis using IoT wearable medical devices. Comput. Methods Biomech. Biomed. Engin. **2021**, 25, 387–411.
4. Diwan, S.; Thakur, G.S.; Sahu, S.K.; Sahu, M.; Swamy, N. Predicting Heart Diseases through Feature Selection and Ensemble Classifiers. J. Phys. Conf. Ser. **2022**, 2273, 012027.
5. AlMohimeed, A.; Saleh, H.; Mostafa, S.; Saad, R.M.A.; Talaat, A.S. Cervical Cancer Diagnosis Using Stacked Ensemble Model and Optimized Feature Selection: An Explainable Artificial Intelligence Approach. Computers **2023**, 12, 200.
6. Miao, L.; Wang,W. Cardiovascular Disease Prediction Based on Soft Voting Ensemble Model. J. Phys. Conf. **2023**, 2504, 012021.
7. Shorewala, V. Early detection of coronary heart disease using ensemble techniques. Inform. Med. Unlocked **2021**, 26, 100655.
8. Jain, V.; Kashyap, K.L. Multilayer Hybrid Ensemble Machine Learning Model for Analysis of COVID-19 Vaccine Sentiments. J. Intell. Fuzzy Syst. **2022**, 43, 6307–6319.
9. Abhishek, H. V. Bhagat and M. Singh, "A Machine Learning Model for the Early Prediction of Cardiovascular Disease in Patients," *2023 Second International Conference on Advances in Computational Intelligence and Communication (ICACIC)*, Puducherry, India, 2023, pp. 1-5, doi: 10.1109/ICACIC59454.2023. 10435210.
10.S. Charkha, A. Zade and P. Charkha, "cardiovascular disease (CVD) Prediction Using Deep Learning Algorithm," *2023 International Conference on Integration of Computational Intelligent System (ICICIS)*, Pune, India, 2023, pp. 1-6, doi: 10.1109/ICICIS56802.2023.10430254.
11.G. M. Shree Raksha, R. Hegde, M. N. Shivani, P. S. Shrinidhi, M. M. Thashwin Monnappa and S. M. Soumyasri, "A Novel Technique for Prediction of Cardiovascular Disease," *2022 IEEE International Conference on Data Science and Information System (ICDSIS)*, Hassan, India, 2022, pp. 1-5, doi: 10.1109/ICDSIS55133.2022.9915934.
12.V. Chowdary B, C. Datta M and R. Senapati, "An Improved Cardiovascular Disease Prediction Model Using Ensembling of Diverse Machine Learning Classifiers," *2023 OITS International Conference on Information Technology (OCIT)*, Raipur, India, 2023, pp. 329-333, doi: 10.1109/OCIT59427.2023.10430692.
13.S. Talapaneni, C. S. Kota, N. Yalagala, R. Nunna and R. Mothukuri, "Enhancing Heart Disease Prediction and Analysis: An Efficient Voting Ensemble model," *2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE)*, Gautam Buddha Nagar, India, 2024, pp. 156-160, doi: 10.1109/ IC3SE62002.2024.10593602.
14.P. Chakraborty, B. K. Sarkar, M. Mundher adnan, S. Srikanth and S. M. Sundaram, "A Post-processing Ensemble Machine Learning Approach for Prediction and Classification of Cardiovascular Disease," *2024 International Conference on Distributed Computing and Optimization Techniques (ICDCOT)*, Bengaluru, India, 2024, pp. 1-4, doi: 10.1109/ICDCOT61034.2024.10515740.
15.S. Mondal, R. Maity, Y. Omo, S. Ghosh, and A. Nag, ‘‘An efficient computational risk prediction model of heart diseases based on dualstage stacked machine learning approaches,’’ IEEE Access, vol. 12, pp. 7255–7270, 2024.
16.M. S. A. Reshan, S. Amin, M. A. Zeb, A. Sulaiman, H. Alshahrani, and A. Shaikh, ‘‘A robust heart disease prediction system using hybrid deep neural networks,’’ IEEE Access, vol. 11, pp. 121574–121591, 2023.

17.Nadiah A. Baghdadi, Sally Mohammed Farghaly Abdelaliem, Amer Malki, brahim Gad, Ashraf Ewis and Elsayed Atlam, "Advanced machine learning techniques for cardiovascular disease early detection and diagnosis," Journal of Big Data (2023) 10:144 https://doi.org/ 0.1186/s40537-023-00817-1.
18.S. Subramani, N. Varshney, M. V. Anand, M. E. M. Soudagar, L. A. Al-Keridis, T. K. Upadhyay, N. Alshammari, M. Saeed, K. Subramanian, K. Anbarasu, and K. Rohini, ''cardiovascular diseases prediction by machine learning incorporation with deep learning,'' Frontiers Med., vol. 10, Apr. 2023, Art. no. 1150933.
19.R. R. Sarra, A. M. Dinar, M. A. Mohammed, and K. H. Abdulkareem, ''Enhanced heart disease prediction based on machine learning and χ2 statistical optimal feature selection model,'' Designs, vol. 6, no. 5, p. 87, Sep. 2022.
20.E. I. Elsedimy, S. M. M. AboHashish, and F. Algarni, ''New cardiovascular disease prediction approach using support vector machine and quantumbehaved particle swarm optimization,'' Multimedia Tools Appl., vol. 83, no. 8, pp. 23901–23928, Aug. 2023.
21.A. Rahim, Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim, and A. W. Muzaffar, ''An integrated machine learning framework for effective prediction of cardiovascular diseases,'' IEEE Access, vol. 9, pp. 106575–106588, 2021. https://doi.org/10.1109/ ACCESS.2021.3098688
22.A. Vinora, E. Lloyds, R. Nancy Deborah, M. S. Anandha Surya, V. Krithik Deivarajan and M. MuthuVignesh, "Heart Disease Prediction using Ensemble Model," 2023 International Conference on Artificial Intelligence and Applications (ICAIA) Alliance Technology Conference (ATCON-1), Bangalore, India, 2023, pp. 1-6, doi: 10.1109/ICAIA57370.2023.10169687.
23.UCI Machine Learning Repository - Heart Disease Cleveland UCI. Accessed Sep. 10, 2024. https://archive.ics.uci.edu/dataset/45/heart+disease.
24.Kaggle Repository - Framingham Heart study dataset. Accessed Sep. 10, 2024. https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset.
25.National Health and Nutrition Examination Survey (NHANES) data are publicly available at https://www.cdc.gov/nchs/nhanes/index.htm. Accessed Sep. 10, 2024.
26.Johnson, A., Bulgarelli, L., Pollard, T., Gow, B., Moody, B., Horng, S., Celi, L. A., & Mark, R. (2024). MIMIC-IV (version 3.0). PhysioNet. https://doi.org/10.13026/hxp0-hg59.
27.Korial, A.E.; Gorial, I.I.; Humaidi, A.J. An Improved Ensemble-Based Cardiovascular Disease Detection System with Chi-Square Feature Selection. Computers 2024, 13, 126. https://doi.org/10.3390/computers13060126