JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A Weighted Performance Voting Ensemble-Based Cardiovascular Disease Detection System Using Accuracy and False Negative Rate Optimisation

[1]T. Jayasudha,
[1]Ph.D Research Scholar,
[1]PG & Research Department of Computer Science,
[1]Sri Sarada College for Women (Autonomous), Salem-16.
jayasudhat.mca@gmail.com
[2]Dr. R. Uma Rani,
[2]Principal (Retired),
[2]Sri Sarada College for Women (Autonomous), Salem-16.
drrumarani66@gmail.com

Abstract

Cardiovascular disease (CVD) is a common health issue in the world as well, and it is noteworthy to detect it at the early stages and save lives. Single machine learning models are not concerned with the complex data of cardiovascular diseases. Ensemble methods are usually more successful as they are a mixture of numerous models to attain more results. However, existing methods are either interested in accuracy or they mix models in naive ways. This research is inspired by the need to possess a superior system that would integrate a strong ensemble model together with an intelligent voting technique. This paper introduces a new Weighted Performance Voting Ensemble-based Cardiovascular Disease Detection System (WPVE-CVDDS) that combines various sophisticated ensemble models, such as Bagging-Boosting Stacking, Soft Voting Ensemble, Bootstrap Ensemble, Augmented Stacking, and Sequential Boosting. They suggest a special weighted performance voting ensemble approach, based on the accuracy rates and false negative rates of models, to improve the reliability of prediction. The pre-processing pipeline uses the missing data imputation method of MissForest, the outlier detection method of the Local Outlier Factor, the class balancing method of K-means with SMOTE, the feature selection method of the mutual information, and the categorical data transformation method of target encoding with cross-validation. The experimental outcomes on real-world cardiovascular data show that the proposed hybrid system is superior and more robust than the traditional techniques and has the potential to be used to effectively screen cardiovascular risks and support decision-making.

Key Words: Cardiovascular Disease, Ensemble classification, Pre-processing, weighted performance voting ensemble approach, Predictive System

1. Introduction

The cardiovascular diseases (CVDs) are the leading cause of death in the world and cause about 30 percent of the world's deaths [1]. The main morbidity of any form of heart disease is the unhealthy diet, absence of exercise, overweight, high systolic BP, high cholesterol level, high level of fasting glucose, high body mass index, tobacco-based smoking, and physical inactivity [2]. A careful daily lifestyle can curb the effects of these risk factors that include eating less salt, eating fruits and vegetables, physical activity, avoiding tobacco use and alcohol, and eventually it might help lower the chances of getting heart disease.[3]. The mortality rate associated with cardiovascular attack is rising uncharacteristically in high-income areas and presents a serious health sector challenge [4]. Early detection of heart disease by use of a prediction model has mostly been suggested to help decrease the mortality rate and enhance the decision-

making towards further prevention and treatment [5]. The implemented prediction model can be applied to the clinical decision support system (CDSS) to assist clinicians in evaluating the risk of heart disease and offering them optimal treatments to contain the risk [6]. Moreover, several studies have also indicated that the use of CDSS can enhance preventative care, clinical decision making and decision quality.

Clinical decision-making, based on machine learning, has recently been introduced into health care. The literature has indicated that machine learning approaches, including logistic regression (LR), support vector machine (SVM), and random forest (RF) [7], have been effectively employed in supporting decision-making in predicting heart disease based on personal data. Several studies have also identified the benefits of hybrid models that can perform well in predicting heart disease, including majority voting of naïve bayes (NB), bayes net (BN), RF, and MLP, two stacked SVMs, and RF with a linear model [8].

2. Literature Review

In a study [8] by S. Mohan, the authors claim 88.7 percent accuracy on a hybrid machine learning algorithm, a Random Forest and Linear Model (HRFLM), on the Cleveland heart disease dataset. This is a hybrid model that is better at prediction than the former standalone models.

Jingyi Zhang et al. [9] suggested an ensemble machine learning model to screen for coronary heart disease through echocardiography and risk factors on the basis of a clinical trial dataset. They initially used Principal Component Analysis (PCA) to reduce dimensionality. A stacking ensemble strategy with numerous classifiers was then used. Their model was found to be accurate in the detection of coronary heart disease at 87.7 percent.

Vaishali M Deshmukh [10] built a prediction system of heart disease based on ensemble techniques with normalization, feature selection with Extra Trees and majority voting with bagging on the Decision Tree, Logistic Regression, ANN, KNN, and Naive Bayes classifiers. Their model was 87.78 per cent accurate on the Cleveland data.

Jeevan Babu Maddala et al. [11] designed a heart failure prediction system with the combination of several datasets, feature extraction in the form of a Random Forest, and balancing in terms of the SMOTE. They applied such classifiers as Random Forest, Gradient Boosting, AdaBoost, Extra Forest, XGBoost and a hybrid model of ET, RF, and XGBoost with a hyperparameter search using Grid Search CV. Their hybrid model had an accuracy of 89.82.

The Cleveland Heart Disease Dataset was used in a study by Bilal Ahmad et al. [12] by applying feature selection methods such as Mutual Information (MI), ANOVA F-test, and Chi-Square test to various machine learning models, such as Neural Network, Logistic Regression, Random Forest, etc. The highest accuracy of 82.3 was obtained with a combination of MI and Neural Network.

3. WPVE-CVDDS Workflow

The pre-processing of the heart disease dataset involves filling in the missing data points with the MissForest imputation, identifying and dropping outliers with the Local Outlier Factor (LOF), and balancing between classes using K-means clustering and SMOTE, finding significant features with the Mutual Information, encoding the categorical variables with the target encoding with cross-validation to avoid data leakage, and scaling all variables with Min-Max normalization. The feature engineering pipeline makes the data well-equipped to develop precise and valid prediction models. Many robust ensemble classifiers were used in my earlier paper [13]. In this study, I enhance the approach by employing these classifiers within a weighted voting mechanism using accuracy and false negatives to predict the final result more accurately. Figure 1 shows the proposed WPVE-CVDDS architecture.

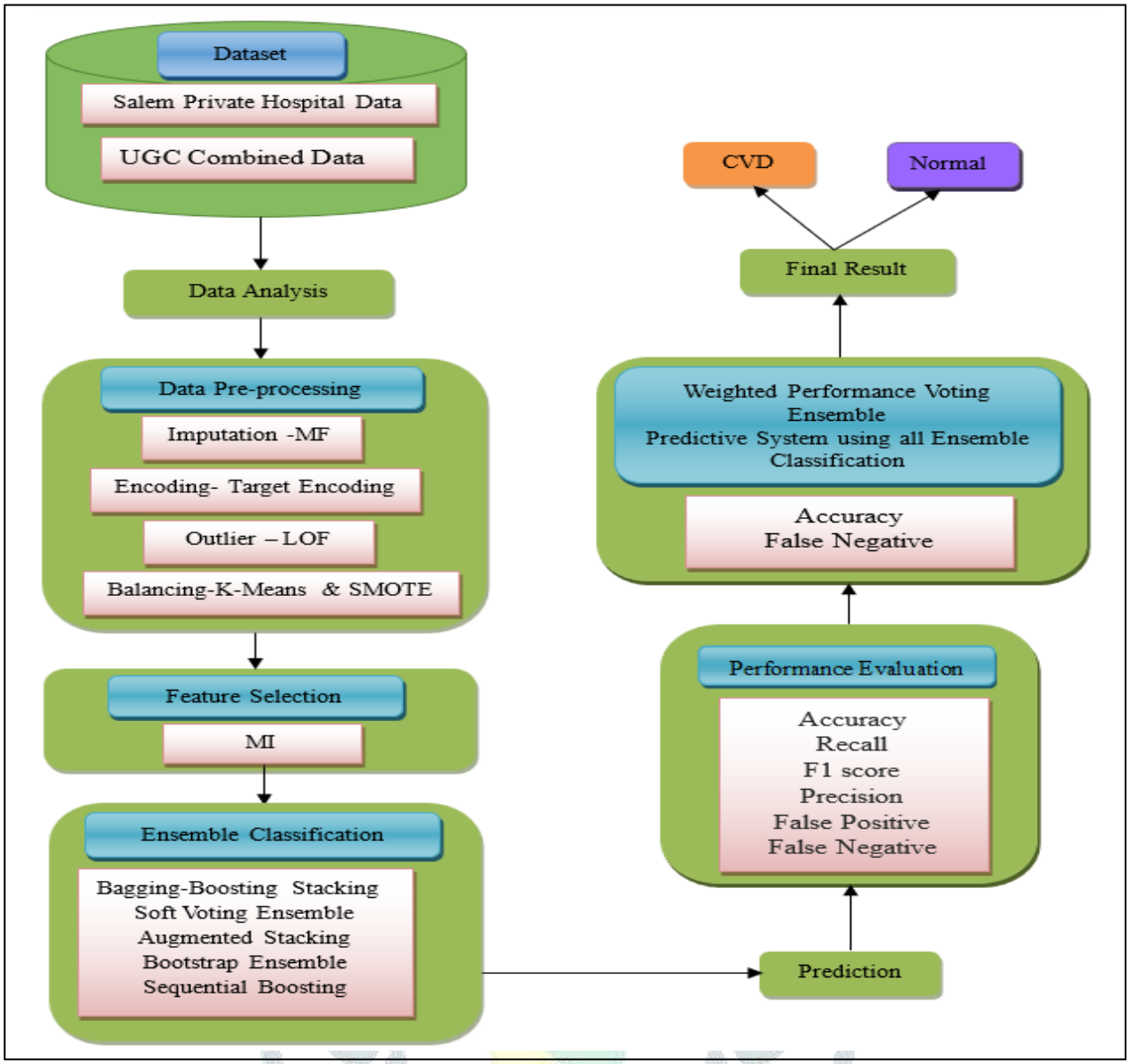

Figure 1: WPVE-CVDDS Architecture

4. Data preparation

Data Collection

For this study, the real-time data were collected from a private hospital in Salem. The data consists of 2,300 entries, one of which is a record of a patient encounter or a case. It has a total of sixteen features, fifteen of which are input features and are used to predict or describe different clinical, demographic or other relevant patient data. The other single feature is the output variable, which constitutes the target or result that is being predicted by the model. This detailed data will give a strong platform to create and confirm predictive models to aid clinical decision-making.

Data Analysis

Exploratory Data Analysis (EDA) can be described as the process of analyzing, summarizing and visualizing a dataset to gain insight into its key characteristics, patterns, anomalies and to test assumptions. It assists data scientists in understanding the structure and relationships of the data prior to using more advanced techniques of analysis or modeling. EDA can be considered an essential initial portion of any data science project to provide an understanding of the data quality, identify latent patterns, and make successful decisions.

Pre-Processing

Imputation

The detection system of cardiovascular diseases exploits a robust preprocessing pipeline beginning with the missing value imputation with MissForest, which is a non-parametric imputation approach using random forests to successively impute missing entries with high precision, especially for mixed-type clinical data.

Outlier Management

In order to deal with data anomalies, the local outlier factor (LOF) is applied to detect outliers and delete samples that are not consistent with the local data density to enhance the robustness of the model by eliminating spurious samples.

Encoding

In the case of categorical feature transformation, target encoding through cross-validation is carried out, where the categorical variables are substituted by the average target value in each category, and it is calculated in the cross-validation folds to avoid data leakage and overfitting.

Normalisation

The step of Min-Max scaling optimizes the features to a range that guarantees the equal distribution of values and better convergence in the next machine learning pipeline.

Class Balancing

To deal with class imbalance, K-means clustering with SMOTE (Synthetic Minority Over-sampling Technique) is used, where K-means is used to split the minority class into smaller parts to create more realistic synthetic samples using SMOTE, which does not increase noise but increases the representative nature of the minority class.

Feature Selection

Mutual Information (MI) is used to select features and to measure the dependency between features and target features, and select the most informative predictors that best help to classify heart diseases.

5. Ensemble Classification

Bagging-Boosting Stacking (BBS) Approach

Likelihoods of predictions based on Extra Trees (ET) using K-fold cross-validation were added to the original data to create enriched training and test sets. LightGBM was then trained as a meta-model, and hyperparameter optimisation of the two models was done using Optuna. Finally, the models were retrained with the best parameters and tested on the test set. Bagging-boosting stacking working architecture is shown in Figure 2.

Figure 2: Bagging-Boosting Stacking Workflow

Soft Voting Ensemble (SVE)

Four classifiers, namely Gaussian Naive Bayes, Logistic Regression, Support Vector Machine and CatBoost, would be integrated into a soft voting ensemble and evaluated using K-fold cross-validation. Optuna was used to optimise hyperparameters for each trial; the best setup would then be used to re-train the final ensemble on the whole training data, test it on the test data and report the performance of the result. The results of the experiment were also visualised through the plotting of every trial. Soft voting ensemble workflow is shown in Figure 3.

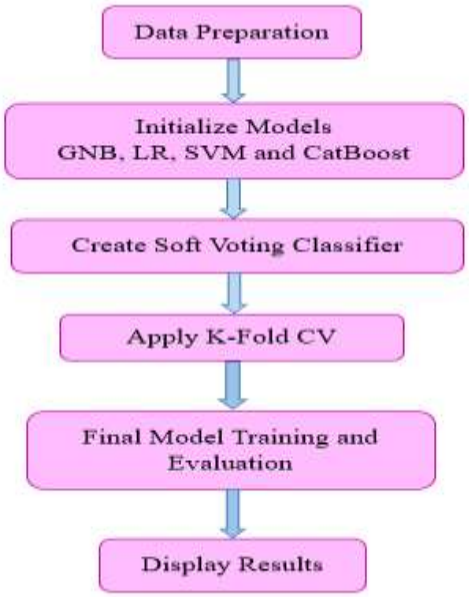

Figure 3: Soft Voting Ensemble Workflow

Augmented Stacking (AS)

There were four base models (LR, ET, SVM, KNN) that were trained using XGBoost as the meta-model. K-fold cross-validation prediction probabilities were added to the original features to generate new datasets. Hyperparameter tuning was conducted with optuna, the models were remodeled using the ideal parameters, and the results of running the models on the test set were provided. The augmented stacking step-by-step process is illustrated in Figure 4.

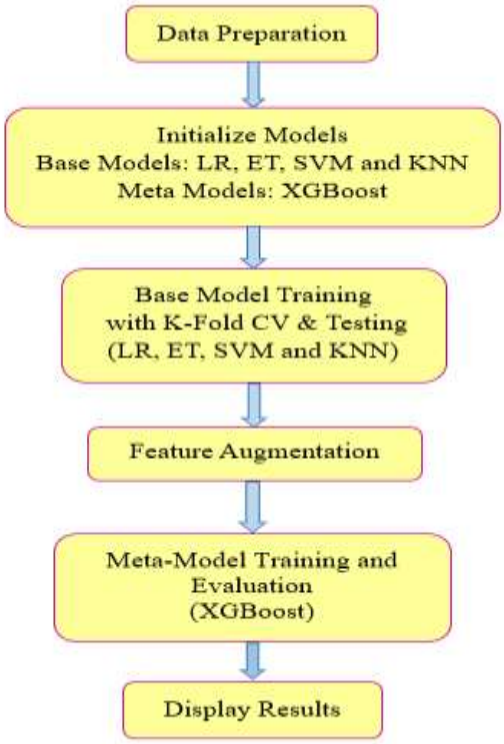

Figure 4: Augmented Stacking Workflow

Bootstrap Ensemble (BE)

Quadratic Discriminant Analysis (QDA), Support Vector Machine (SVM), and K-Nearest Neighbours (KNN) models were bagged, and a Random Forest (RF) classifier was included directly because of its inherent bagging support. These were combined into a soft-voting ensemble, whose hyperparameters were optimized using Optuna. The accuracy of the ensemble was first tested using K-fold cross-validation, followed by re-training using the complete training set, testing the model using the test set and illustrating its performance using visual plots. The Bootstrap ensemble workflow is shown in Figure 5.

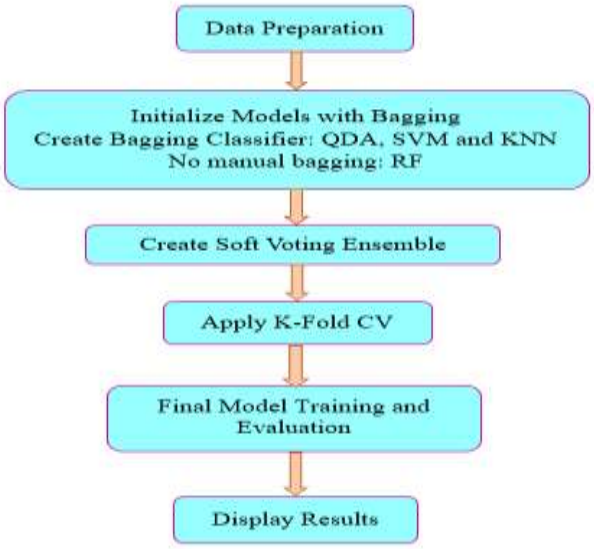

Figure 5: Bootstrap Ensemble Workflow

Sequential Boosting (SB)

An objective function was developed in Optuna to maximize the parameters of the model by means of K-fold cross-validation, in which the models were trained consecutively by applying error-adjusted sample weights and evaluated on the held-out folds. Optuna found the best hyperparameters based on the average accuracy achieved at the trials. This last model was then retrained using the entire training set and tested using the test set, and the performance was represented visually. The sequential boosting step-by-step process is illustrated in Figure 6.

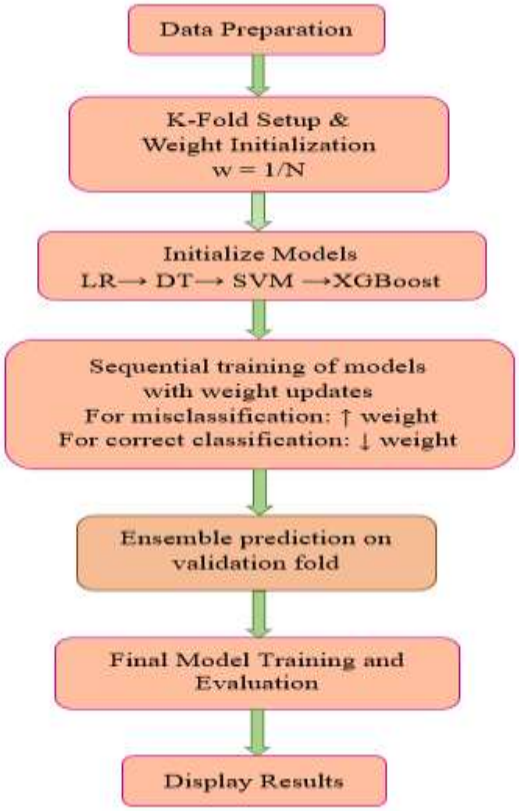

Figure 6: Sequential Boosting Workflow-

6. Implementation

EDA

A summary of data is a brief description of data that brings out the principal features of the data. It generally involves some of the important statistics of the mean, median, mode, range, variance, and standard deviation. These statistics are used to know the central tendency, spread and general distribution of the data. Data summaries can give a brief understanding of how the data is structured, what trends or anomalies are present in the data, and a starting point for further analysis. Figure 7 shows a statistical summary of the data.

	count	mean	std	min	25%	50%	75%	max
Gender	2300.0	0.625652	0.484059	0.0	0.000000	1.000000	1.000000	1.0
Age	2300.0	53.544348	19.064033	2.0	41.000000	55.000000	66.000000	100.0
Blood Pressure	2300.0	141.890522	34.345925	80.0	120.000000	130.000000	170.000000	297.0
Pulse Rate	2300.0	89.558696	17.492247	13.0	80.000000	86.000000	98.000000	197.0
Temperature	2300.0	98.586304	0.802535	97.0	98.400000	98.400000	98.600000	104.0
Respiratory Rate	2300.0	22.156522	2.806013	12.0	20.000000	22.000000	22.000000	86.0
SpO_2	2300.0	97.056087	3.063896	56.0	97.000000	98.000000	98.000000	100.0
Complaints	2300.0	0.360000	0.340689	0.0	0.082569	0.297468	0.640625	1.0
Obesity	2300.0	0.038261	0.191867	0.0	0.000000	0.000000	0.000000	1.0
Anemia	2300.0	0.061304	0.239940	0.0	0.000000	0.000000	0.000000	1.0
Cholesterol	2300.0	227.517826	61.449719	13.0	189.000000	210.000000	276.000000	978.0
Glucose	2300.0	123.607826	42.451980	62.0	98.000000	110.000000	133.000000	450.0
Asthma	2300.0	0.007391	0.085673	0.0	0.000000	0.000000	0.000000	1.0
Hypertension	2300.0	0.359565	0.479977	0.0	0.000000	0.000000	1.000000	1.0
Diabetes	2300.0	0.305217	0.460600	0.0	0.000000	0.000000	1.000000	1.0
CVD	2300.0	0.360000	0.480104	0.0	0.000000	0.000000	1.000000	1.0

Figure 7: Statistical Summary of Data

Data distribution is the method of distributing the data values among the various possible values or intervals. It indicates the frequency of different data points in a dataset. Knowledge of the data distribution can be used to determine trends, identify skewness, symmetry and outliers, and is used to make decisions about the suitable analysis or modeling methods. Figure 8 denotes the distribution of numerical data. Figure 9 denotes the distribution of categorical data.

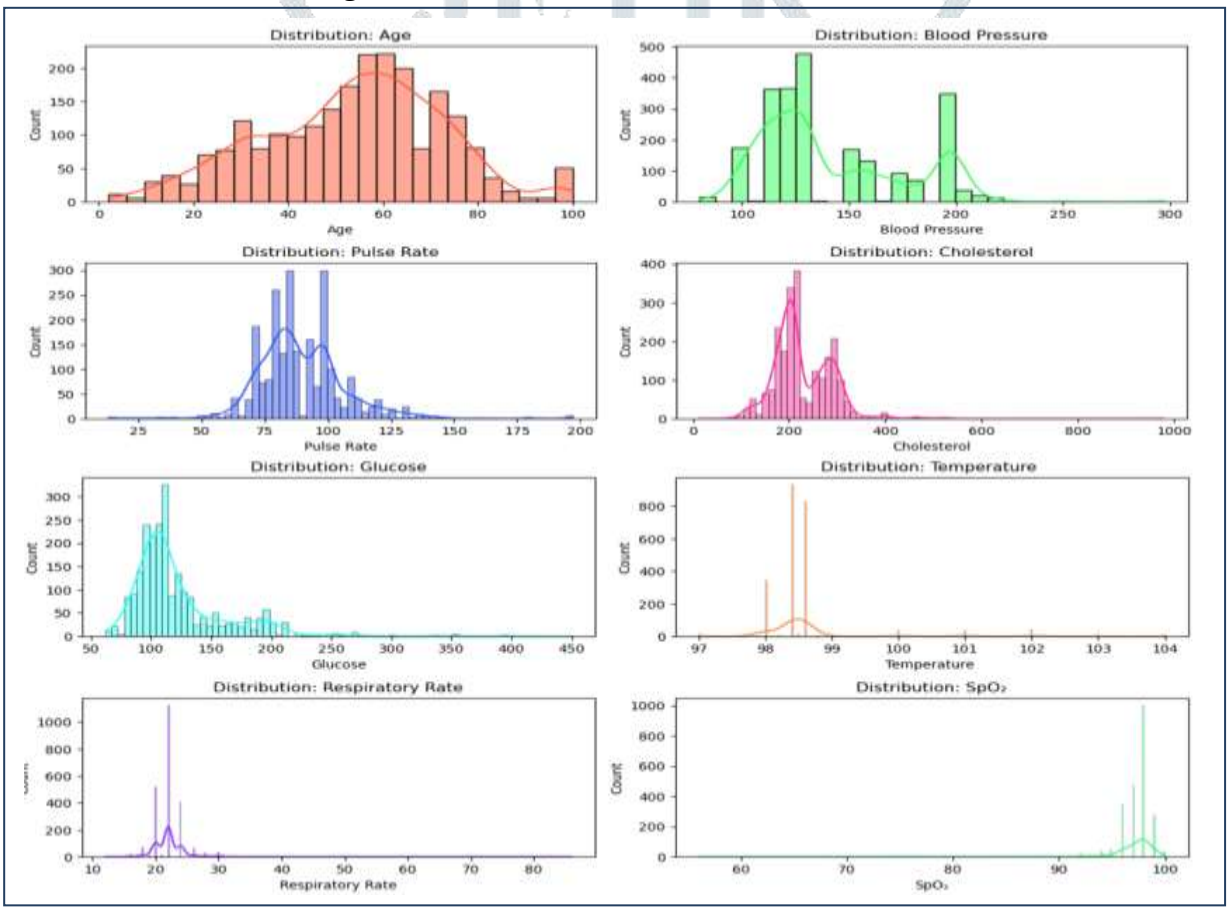

Figure 8: Numerical data Distribution

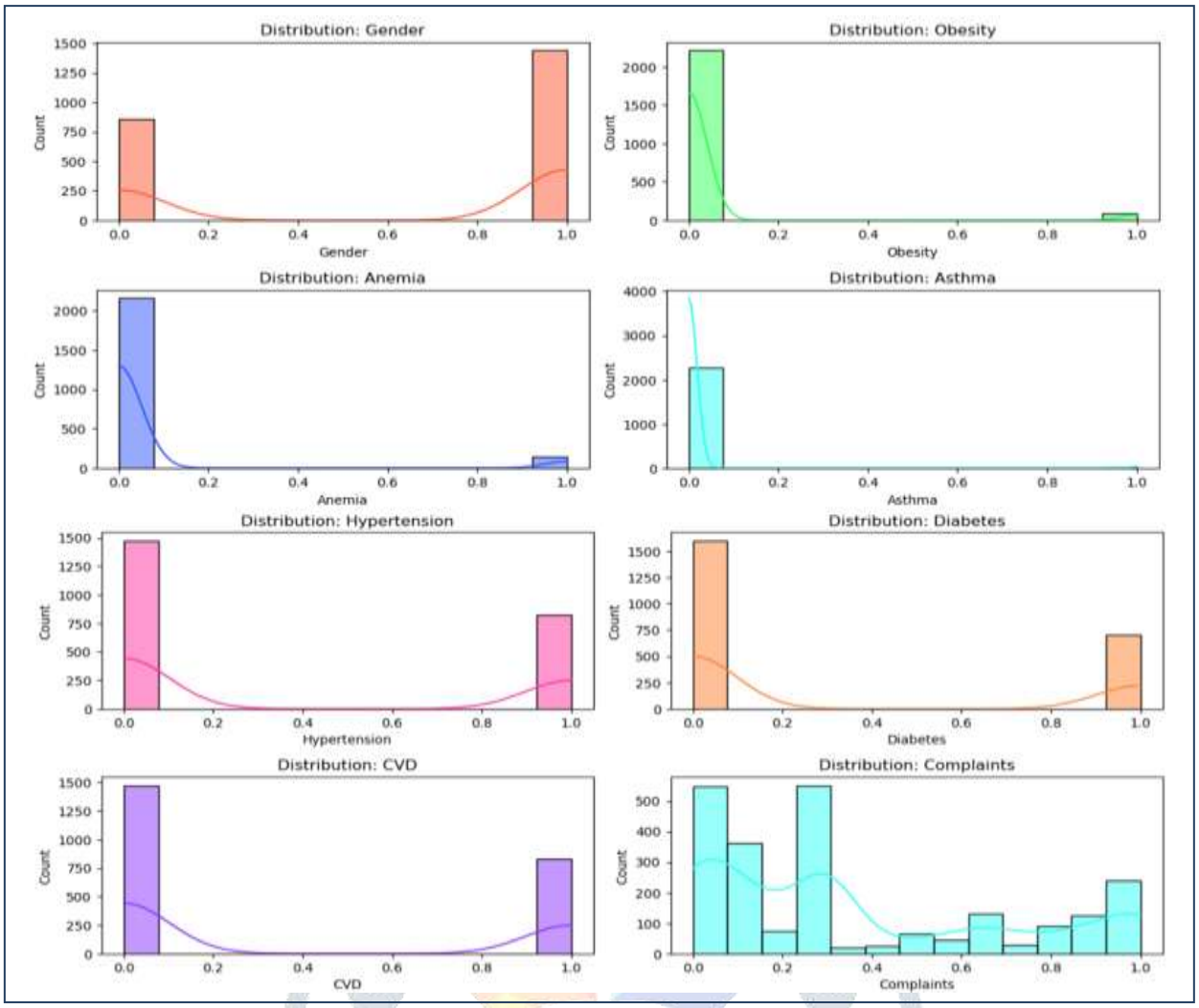

Figure 9: Categorical Data Distribution

Feature Selection -Mutual Information

Figure 10 displays the mutual information scores of all features. The best 8 features will be used for classification.

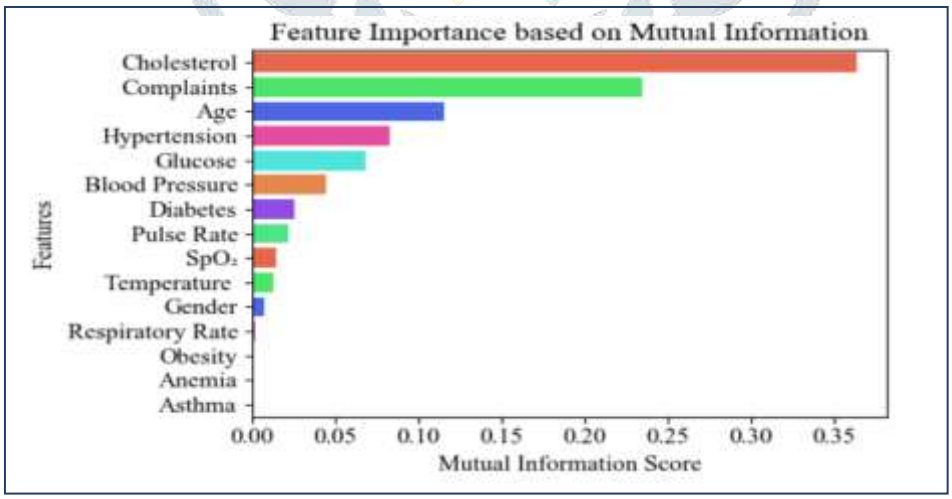

Figure 10: Mutual Information Scores

Classification Results

The performance measures of each model are outlined in Figure 11. The ensemble classifiers that performed the best were augmented stacking with 94.7% accuracy, and sequential boosting with 94.0%. Soft voting ensemble got an accuracy of 93.5% bootstrap ensemble got an accuracy of 92, and bagging-boosting stacking gave an accuracy of 91.0%.

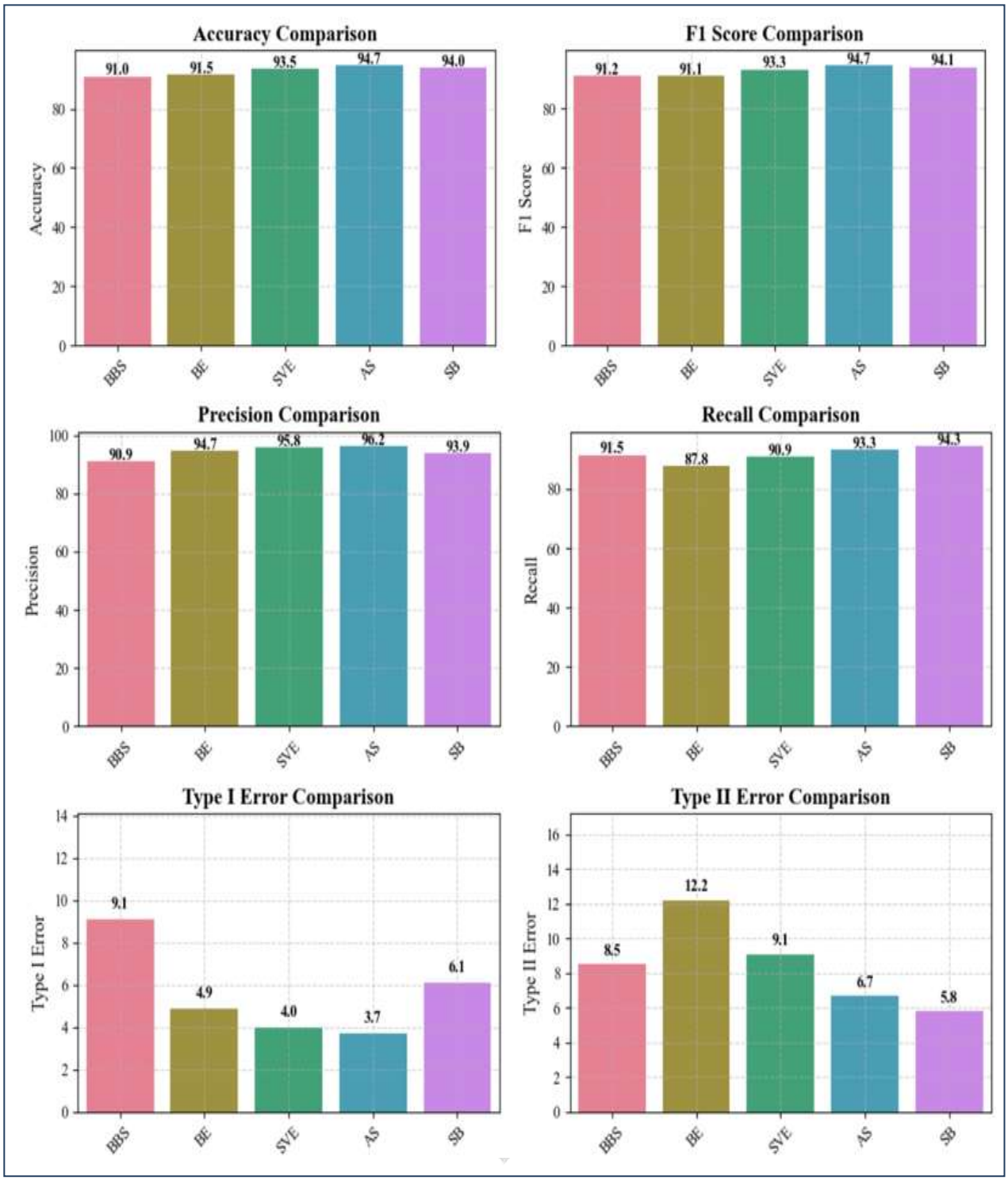

Figure 11: Ensemble Model Performance

Predictive System

A new weighted performance voting ensemble approach combines the projection of multiple models by maximizing accuracy and reducing false negatives to yield a robust result. This technique, in contrast to conventional voting techniques, gives every model a weight to balance between its overall accuracy and the ability to minimize false negatives, which are important in most uses, such as in healthcare. Through the process, weighted soft voting is used to combine the predictions of each of the models with the weights calculated based on measures of the two objectives. Such a two-fold concentration will make the ensemble not just achieve high overall correctness, but also minimize costly errors, which have negative effects on results. This new system is more predictive because it meets several objectives of performance at the same time. It is a mechanism that gives weights to the prediction of each model according to these objectives, thus improving the overall prediction accuracy and minimizing critical false negatives, often in the form of a weighted sum (Equation 1):

$$\text{Final Score} = w_1 * \text{Accuracy} + w_2 * (1\text{-False Negative Rate}) \quad (1)$$

Where w 1 and w 2 are weights that represent the significance of the accuracy and false-negative minimization.

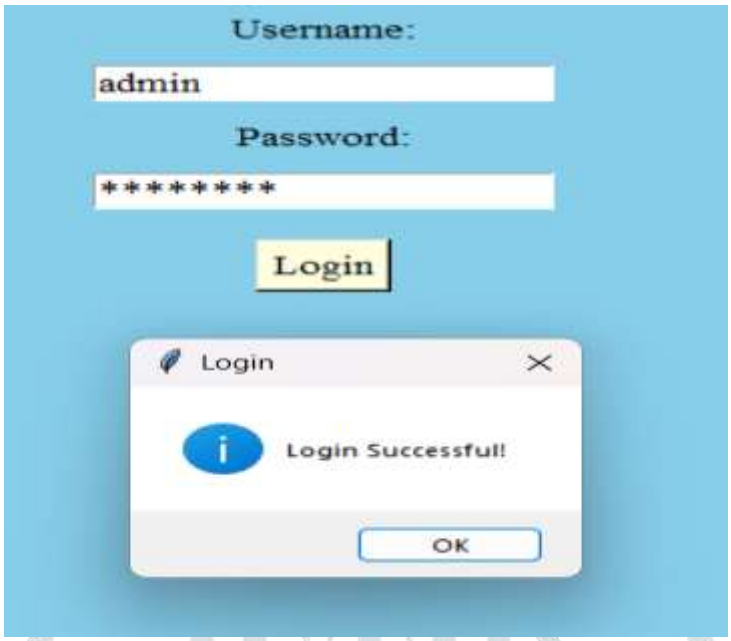

Figure 12: Login Page of WPVE-CVDDS

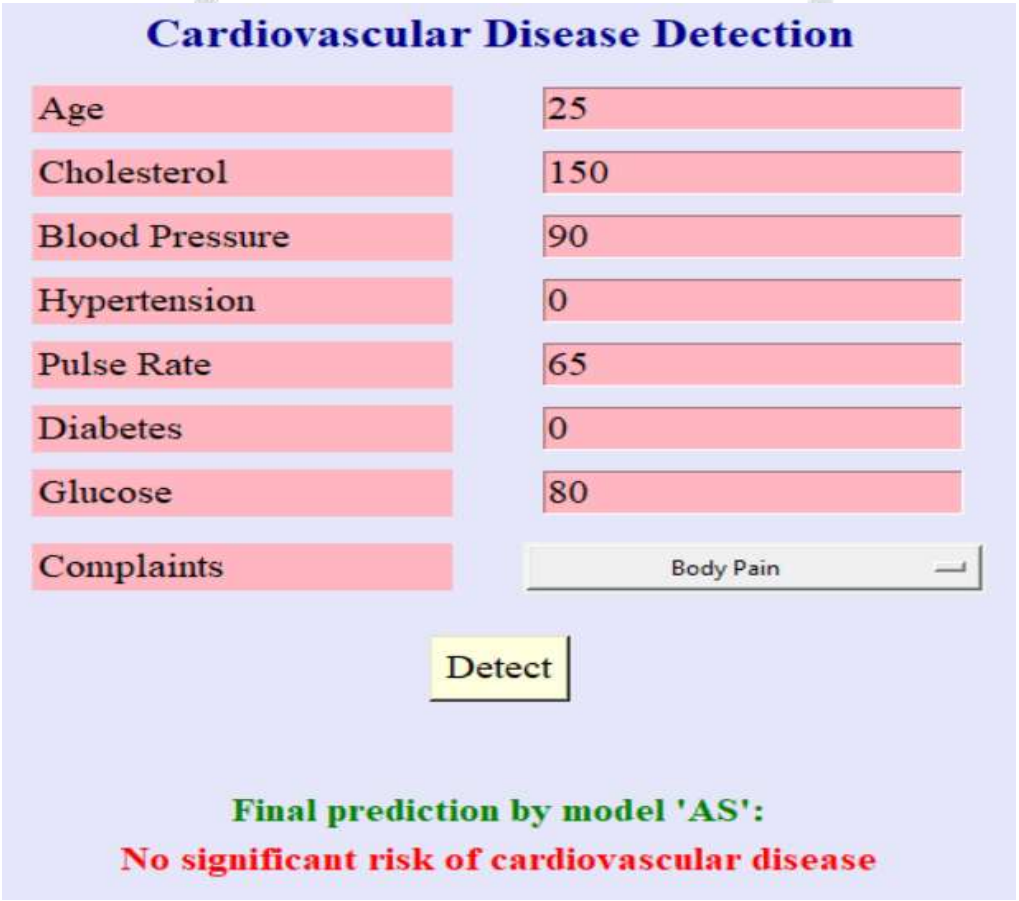

Figure 13: Result Page of WPVE-CVDDS

Figure 12 shows the login page of the WPVE-CVDDS approach. Figure 13 indicates that the patient, aged 25, has a cholesterol level, which is in the healthy range (150), the normal blood pressure (90), has no history of hypertension, a pulse rate of 65, no diabetes, a regular glucose level (80), and body pain complaints. The patient does not have cardiovascular disease. The AS model anticipated this outcome, as it scored the highest performance compared to all the other models that were tested.

7. Conclusion

The current research introduces a new system of cardiovascular disease detection, which integrates several powerful models of ensembles, such as Bagging-Boosting Stacking, Soft Voting, Bootstrap Ensemble, Augmented Stacking, and Sequential Boosting, inside a multi-objective voting framework. In contrast to prior research that uses voting on simple models or optimized ensembles individually, our methodology uses a combination of different robust ensemble models and uses a voting mechanism that combines the forecasts of multiple models by using two objectives, such as accuracy and false negatives, to conclude on the outcome. The multi-stage methodology and feature engineering pipeline guarantee enhanced prediction results, showing a remarkable progress in the methods of the ensemble machine learning applied to the cardiovascular disease prediction. The practice of combining several strong models of the ensemble with voting in the final prediction can be regarded as innovative, since this multi-model voting is not similar to the process of voting on individual models, and demonstrates better performance.

Reference

[1] World Health Organization. (2017). *Cardiovascular Diseases (CVDs)*. [Online]. Available: https://www.who.int/health-topics/cardiovasculardiseases/

[2] Willinger L, Brudy L, Meyer M, Oberhoffer-Fritz R, Ewert P, Müller J. 2021. Prognostic value of non-acute high-sensitive troponin-T for cardiovascular morbidity and mortality in adults with congenital heart disease: A systematic review. Journal of Cardiology, 78(3): 206-212

[3] World Health Organization. (2017). *Cardiovascular Diseases (CVDs)*. [Online]. Available: https://www.who.int/news-room/factsheets/ detail/cardiovascular-diseases-(cvds)

[4] Faizal ASM, Thevarajah TM, Khor SM, Chang SW. 2021. A review of risk prediction models in cardiovascular disease: conventional approach vs. artificial intelligence approach. Computer Methods and Programs in Biomedicine, 207: 106190

[5] G.-M. Park and Y.-H. Kim, ``Model for predicting cardiovascular disease: Insights from a Korean cardiovascular risk model," *Pulse*, vol. 3, no. 2, pp. 153_157, 2015, doi: 10.1159/000438683.

[6] G. J. Njie, K. K. Proia, A. B. Thota, R. K. C. Finnie, D. P. Hopkins, S. M. Banks, D. B. Callahan, N. P. Pronk, K. J. Rask, D. T. Lackland, and T. E. Kottke, ``Clinical decision support systems and prevention," *Amer. J. Preventive Med.*, vol. 49, no. 5, pp. 784_795, Nov. 2015, doi: 10.1016/j.amepre.2015.04.006.

[7] A. Gupta, R. Kumar, H. S. Arora, and B. Raman, ``MIFH: A machine intelligence framework for heart disease diagnosis," *IEEE Access*, vol. 8, pp. 14659_14674, 2020, doi: 10.1109/ACCESS.2019. 2962755.

[8] S. Mohan, C. Thirumalai, and G. Srivastava, ``Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542_81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[9] Jingyi Zhang et al., "Ensemble machine learning approach for screening of coronary heart disease based on echocardiography and risk factors", BMC Med Inform Decis Mak (2021) 21:187 https://doi.org/10.1186/s12911-021-01535-5

[10] Vaishali M Deshmukh, "*Heart Disease Prediction using Ensemble Methods*", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-3, September 2019.

[11] Jeevan Babu Maddala, Bhargav Reddy Modugulla, "Heart Failure Prediction Using Machine Learning", 2024, DOI Link: https://doi.org/10.22214/ijraset.2024.59236

[12] Bilal Ahmad et al., "Feature selection strategies for optimized heart disease diagnosis using ML and DL models", 2025, https://doi.org/10.48550/arXiv.2503.16577.

[13] T. Jayasudha, Dr R. Uma Rani, "An Innovative Machine Learning Framework for Cardiovascular Disease Detection Incorporating Feature Selection, Cross-Validation, and Ensemble Classification", Submitted.