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 Majority Voting Ensemble Framework for Cardiovascular Disease Risk Assessment

S. Peruvazhuthi[1], K. Praveen[2]

[1]M.E, Department of Electronics and Communication Engineering, College of Engineering, Guindy, Anna University, Chennai, Tamil Nadu 600025, India

[2]Assistant professor, Department of Electronics and Communication Engineering, College of Engineering, Guindy, Anna University, Chennai, Tamil Nadu 600025, India

Abstract: The global mortality rate associated with cardiovascular disease (CVD) continues to rise each year. Timely prediction of the disease has the potential to save millions of lives. Machine Learning (ML) algorithms utilized for these tasks are capable of recognizing subtle patterns and risk factors that might not be evident to medical practitioners. However, existing ML methodologies lack the ability to deliver enhanced accuracy and performance. An ensemble technique, the Weighted Majority Voting Ensemble (WMVE) classifier, is implemented to improve the performance of CVD prediction. The WMVE classifier undergoes training and testing using the "Sathvi" dataset, a composite of the Cleveland, Hungarian, Long Beach, and Switzerland datasets. Pearson's correlation technique was utilized for feature selection to remove highly correlated feature. Employing Multilayer Perceptron (MLP), Extreme Gradient Boosting (XGBoost), CatBoost, and Logistic Regression (LR) as the base classifiers, the WMVE model has achieved an accuracy rate of 88.78% and an Area Under the Receiver Operating Characteristic curve (AUROC) of 0.899, outperforming all base classifiers.

***Keywords*- cardiovascular disease, machine learning algorithms, weighted majority voting, ensemble technique, feature selection.**

I. INTRODUCTION

Accounting for approximately 32% of global deaths, the World Health Organization (WHO) assessed those 1.79 crore individuals expired due to cardiovascular diseases (CVDs) in 2019. This trend of increasing CVD-related casualties has persisted in the subsequent years as well [1]. The term 'cardiovascular disease' (CVD) denotes a category of illnesses affecting the heart or blood vessels. Coronary artery disease (CAD) is the most prevalent form of heart disease among adults worldwide. It is commonly linked to the buildup of lipid deposits within the arteries, a condition known as atherosclerosis. This condition impacts the blood flow to the heart, and reduced blood flow may lead to a cardiac event. CVD risk is significantly influenced by hypertension, elevated blood cholesterol levels, and usage of tobacco. Nearly half of the United States population (47%) possesses at least one of these three risk factors. Adopting a healthy lifestyle can reduce the likelihood of developing CVD [2,3].

D. Stojanov et al [4] investigated 167 patients admitted to Villa Scassi Hospital in Genoa, Italy, focusing on predicting coronary artery disease versus cardiac failure. They analyzed nine biochemical parameters, including Triglycerides, Aspartate Aminotransferase, C-reactive protein, Alanine Transaminase, High-sensitive Cardiac Troponin I, Hemoglobin, High-density Lipoprotein, Low-density Lipoprotein, and Serum Creatinine. Through feature selection, they identified Serum Creatinine, High-sensitive Cardiac Troponin I, Aspartate Aminotransferase, Hemoglobin, and C-reactive protein as crucial parameters for diagnosing coronary artery disease against cardiac failure. The logistic regression model was utilized for prediction, yielding an AUROC of 0.805. Due to the limited dataset containing only 167 patient records, a lower AUROC is observed. The work presented by I. Tougui et al [5] utilized the Cleveland heart disease dataset, consisting 303 instances with 14 attributes. The authors used six machine learning classification models, namely Naive Bayes, Random Forest, Logistic Regression, Artificial Neural Network, K Nearest Neighbors, and Support Vector Machine. The optimal model, achieving an accuracy of 85.86%, was the Artificial Neural Network. Further improvement in accuracy is suggested through hyperparameter tuning.

II. RELATED WORKS

A. Ghasemieh et al [6] employed a proprietary dataset from the Massachusetts Institute of Technology (MIT) Laboratory for Computational Physiology. Their approach involved a stacking ensemble, with base models including Support Vector Machines, Random Forest, K-Nearest Neighbors, Decision Tree, Extreme Gradient Boosting, and Logistic Regression, while Extreme Gradient Boosting served as the meta-learner. The stacking ensemble achieved an accuracy of 88%. However, its hierarchical structure introduces higher computational complexity, posing a potential drawback in clinical settings. B. A. Tama et al [7] introduced a two-tier ensemble technique, combining the gradient boosting machine, random forest, and XGBoost in a stacked structure. Feature selection was implemented using the particle swarm optimization (PSO) method. The classification model underwent training and testing on diverse heart disease datasets, including Cleveland, Z-Alizadeh Sani, Hungarian, and Statlog. The resulting classification model achieved the highest accuracy of 85.71%, surpassing the performance of all its individual base classifiers.

K. V. V. Reddy et al [8] employed ten classifier algorithms on the Cleveland heart disease dataset. Optimal attributes were selected through attribute evaluators, including ReliefF attribute evaluation, feature selection based on correlation, and chi-squared attribute evaluation. When using optimal attributes from the chi-squared attribute evaluator, the Sequential Minimal Optimization (SMO) classifier obtained an accuracy score of 86.468%, and 85.148% with complete attributes. The authors suggested that combining multiple datasets could provide more instances for optimal feature selection, leading to an improved predictive model for cardiovascular disease (CVD). The work presented by C. Gazeloğlu et al [9] employed eighteen machine learning algorithms and three feature selection methods namely feature selection based on correlation, chi-square feature selection, and fuzzy rough set for the Cleveland heart disease dataset. The Radial Basis Function (RBF) Network algorithm achieved an accuracy of 81.188% when utilizing chi-square feature selection. Naive Bayes (NB) resulted in an accuracy of 84.818% with feature selection based on correlation, while the support vector machine (PolyKernel) algorithm achieved an accuracy score of 85.148% without feature selection.

C. B. C. Latha at al [10] explored multiple ensemble techniques, including bagging, majority voting, boosting, and stacking on the Cleveland heart disease dataset. They selected classifiers such as Naive Bayes (NB), Bayes Net (BN), C4.5, Projective Adaptive Resonance Theory (PART), Multilayer perceptron (MLP), and Random Forest (RF). The majority voting ensemble yielded the highest accuracy of 85.48%, with Naive Bayes, Multilayer perceptron, Bayes Net, and Random Forest as the base classifiers. The study observed enhanced performance across all ensemble techniques when integrating feature selection methods.

The work presented by D. Ananey-Obiri et al [11] utilized the Gaussian Naive Bayes, decision tree classifier, and logistic regression for the Cleveland heart disease dataset. A feature reduction process was implemented by the single value decomposition method, reducing the number of features from 13 to 4. Both Gaussian Naive Bayes and logistic regression achieved the maximum accuracy score of 82.75% for CVD prediction. Fig. 1 depicts the heart disease prediction method.

Several studies have been conducted on various ensemble techniques; however, there is a lack of research on training a simple and computationally efficient ensemble classifier with a dataset containing more instances and without missing values. This study aims to implement a weighted majority voting ensemble (WMVE) classifier with an optimal and diverse selection of base classifiers, and the hyperparameters of these classifiers will be finely tuned to improve the performance of the WMVE in predicting the likelihood of CVD.

III. METHODOLOGY

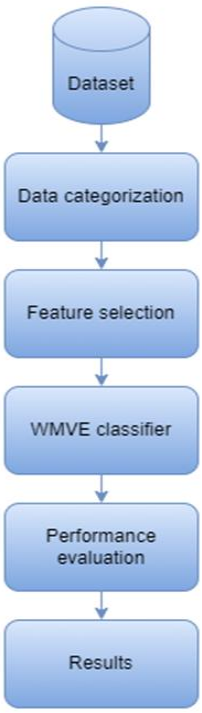

Fig. 1. Heart disease prediction method.

3.1 Dataset

The WMVE classifier was developed using the 'Sathvi' dataset, which is a combination of the Cleveland, Hungarian, Long Beach, and Switzerland datasets available in the UCI Machine Learning Repository [13]. The dataset comprised 531 instances with 12 features and is devoid of missing values. The 'Sathvi' dataset is derived from a Combined or Hybrid dataset, which consists of 920 instances with 14 features, and exhibits 621 missing values. Due to the presence of more than 50% missing values in the 'ca' and 'thal' features, these attributes were discarded from the hybrid dataset. Additionally, row instances with missing values were excluded from the dataset. The dataset produced is denoted as the 'Sathvi' dataset, and both the Sathvi and Hybrid datasets can be found in the supplemental file provided in reference [12]. The visualization of the 'Sathvi' dataset is presented in Fig. 2. It displays the distribution of target values, where 0 indicates instances not at risk of CVD, and 1, 2, 3, and 4 indicate varying levels of CVD risk.

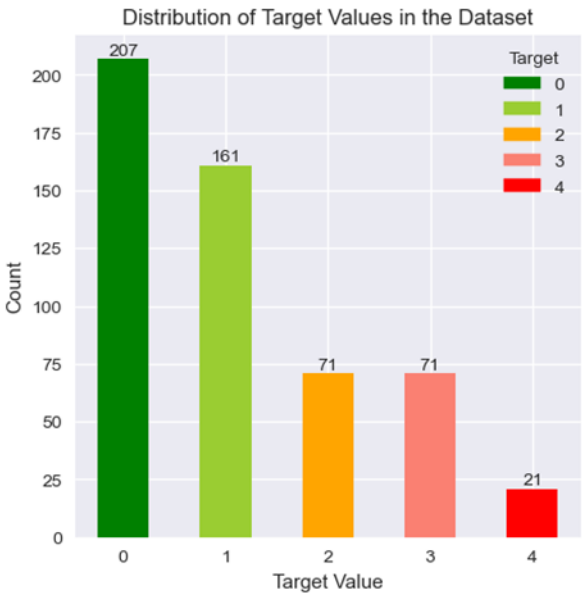

Fig. 2. Graphical representation of the 'Sathvi' dataset.

3.2 Data categorization

Binary classification was employed on the 'Sathvi' dataset, where the target attribute values ranging from 1 to 4 were converted to 1, representing the positive class, indicating individuals at risk of CVD, and 0, representing the negative class, indicating those without CVD risk. Fig. 3 represents the binary classification analysis. It illustrates the distribution of heart disease presence (1) and absence (0), showing that 324 patients are at risk of CVD, while 207 patients are not.

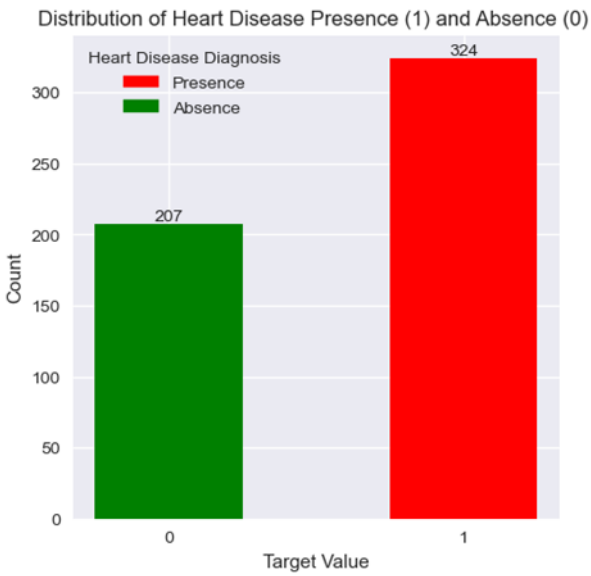

Fig.3. Analysis of binary classification.

3.3 Feature selection

When training the model, two or more features can convey essentially the same information to the model, resulting in redundant information that prevents generalization rather than enhancing model complexity. To mitigate this redundancy and achieve an optimal number of features, feature selection technique known as the Pearson correlation coefficient $R(X,Y)$ is employed [14]. This coefficient assesses the correlation between each pair of attributes, encompassing values within the range of -1 to 1. -1 signifies an inverse linear correlation, 0 denotes no linear correlation, and 1 signifies a linear correlation. Eq. (1) defines the Pearson correlation coefficient $R(X,Y)$. The Pearson correlation heatmap in Fig. 4 illustrates a considerable correlation between the "slope" and "oldpeak" features. To prevent redundant information, one of these features, namely "slope" is discarded. Subsequently, the dataset without the "slope" feature is employed for training and testing the classifier.

Consider a dataset (X,Y) with a sample size of n. $R(X,Y)$ represents the Pearson correlation coefficient between X and Y, $Cov(X,Y)$ is the covariance between X and Y, $\sqrt{D(X)}$ and $\sqrt{D(Y)}$ denote the standard deviations of X and Y respectively, and $\bar{x}$ and $\bar{y}$ represent the means of X and Y respectively.

$$R(X,Y) = \frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$$

$$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (1)$$

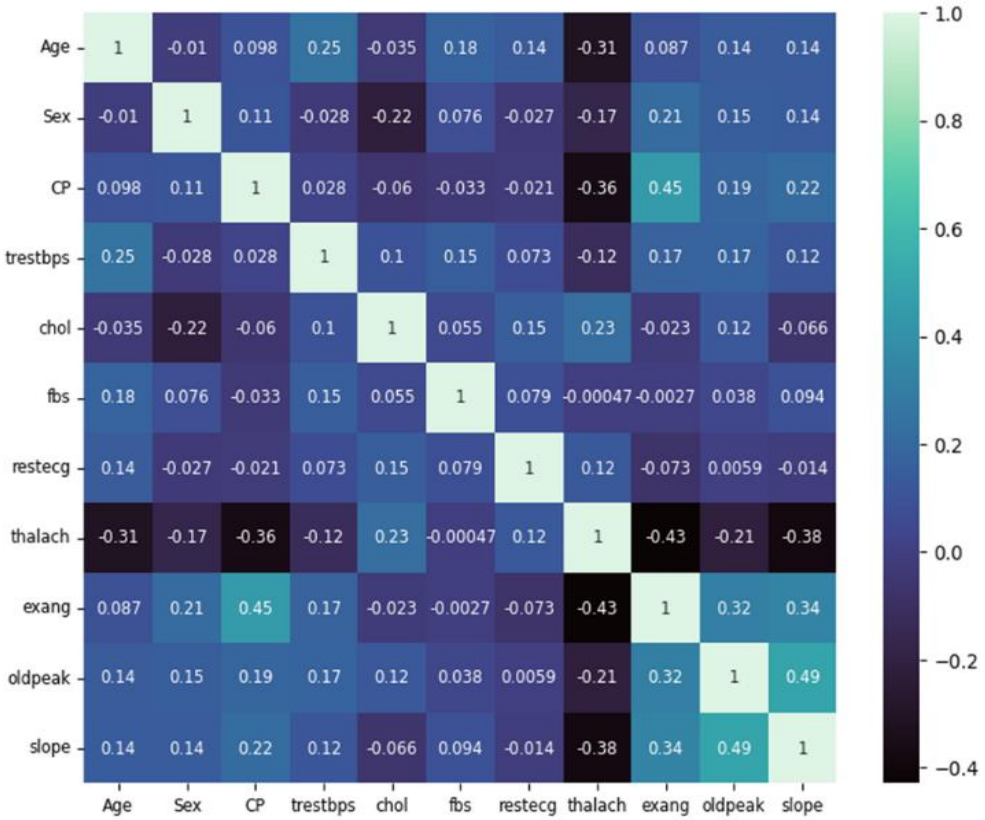

Fig. 4. Pearson's correlation coefficient heatmap.

IV. Proposed ensemble model

4.1 The Weighted Majority Voting Ensemble (WMVE) classifier

The Weighted Majority Voting Ensemble (WMVE) classifier enhances the performance and robustness of CVD predictive model through the integration of predictions from multiple algorithms. The process underlying the WMVE is depicted in Fig. 5. It employs Logistic Regression, Extreme Gradient Boosting, CatBoost, and Multilayer Perceptron as its base classifiers. Predictions from each classifier are combined through majority or hard voting, wherein the class label with the maximum number of votes from all the base classifiers is considered the final prediction for the WMVE classifier. The weights for each classifier are assigned, with a lower weight set at 1 and a higher weight at 2. This configuration imparts twice the voting power to the classifier compared to the other classifiers. Assigning a higher weight to the Multilayer Perceptron (MLP) considerably improves the performance of the WMVE classifier [15] compared to assigning a higher weight to other classifiers.

Let $d_{i,j}$ symbolizes the decision of the classifier C_i regarding class l_j, taking the value 1 if C_i selects l_j and 0 otherwise. In the process of weighted majority voting, the combination of decisions from all classifiers will lead to the choice of class l_j if

$$\sum_{i=1}^{I} w_i d_{i,j} \geq \max_{j=1}\left\{\sum_{i=1}^{I} w_i d_{i,j}\right\} \tag{2}$$

i.e., in Eq. (2) if the cumulative weighted vote assigned to class l_j exceeds or equals the total vote received by any other class.

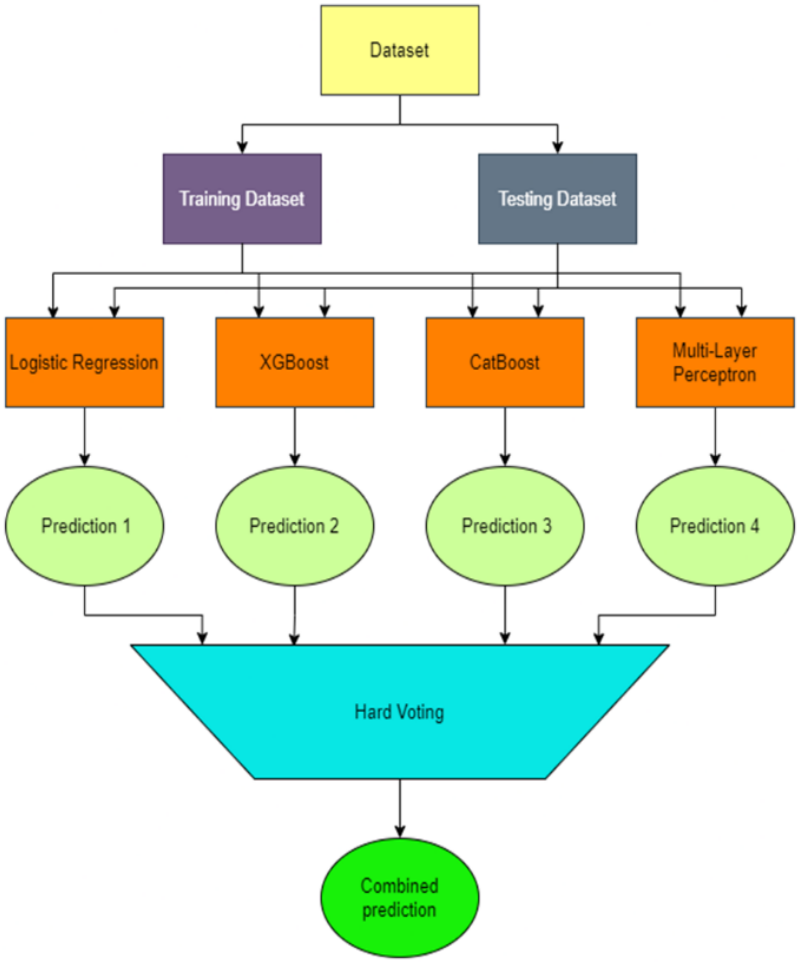

Fig. 5. Operation of WMVE classifier.

4.2 Base classifiers

4.2.1 Logistic regression

Binary classification problems are commonly addressed using logistic regression [16]. It employs the logistic or sigmoid function, as indicated in Eq. (3) and Eq. (4), to estimate the probability of an instance belonging to one of the two classes. The resulting probability values fall within the range of 0 to 1. The threshold serves as a decision boundary, determining whether the predicted probability corresponds to class 0 or class 1.

$$P(X;\beta_0,\beta) = Pr(Y=1|X=x;\ \beta_0,\beta) = \frac{e^{\beta_0+x^T\beta}}{1+e^{\beta_0+x^T\beta}} \tag{3}$$

$$1-P(X;\beta_0,\beta) = Pr(Y=0|X=x;\ \beta_0,\beta) = \frac{1}{1+e^{\beta_0+x^T\beta}} \tag{4}$$

$\beta=(\beta_0,\beta_1,\dots,\beta_n)$, denote the regression coefficients in Eq. (3) and Eq. (4); $x=(x_1,x_2,\dots,x_n)^T$; $X=(X_1,X_2,\dots,X_n)^T$, where $X_1,X_2,\dots,X_n$ denote the n independent variables, Y be the dependent variable.

4.2.2 Extreme gradient boosting

XGBoost, abbreviated from Extreme Gradient Boosting, is based on a distributed gradient-boosted decision tree (GBDT) [17]. It constructs a sequence of weak learners, typically decision trees, wherein each subsequent tree rectifies the errors of its predecessors. Introducing regularization in its cost function, XGBoost aims to mitigate overfitting, thereby enhancing generalization. Eq. (5) defines the loss function of the XGBoost algorithm.

(5)

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i{}^{(t-1)} + f_t(X_i)) + \Omega(f_t)$$
$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \quad (6)$$

Here, $l(y_i, \hat{y}_i{}^{(t-1)} + f_t(X_i))$ denotes the prediction residual of the ith sample at the tth iteration, X_i denotes the ith sample, $\Omega(f_t)$ in Eq. (6) denotes the regularization term, ω_j^2 is the leaf node score, T is the count of leaf nodes, γ denotes the coefficient, and λ is the coefficient for the sum of weights [18].

4.2.3 CatBoost

CatBoost, short for Categorical Boosting, is based on a gradient boosting framework. Its specialized design enables efficient handling of categorical features without the requirement for preprocessing like one-hot encoding. CatBoost yields great results with default parameters, minimizing the necessity for extensive hyperparameter tuning. During the initial split calculation, quantization is utilized for each numerical feature to identify potential data-splitting strategies [19]. The transformation of categorical features into numerical features is indicated in Eq. (7)

Consider a random ordering of the sample, denoted as $\sigma = (\sigma_0, \sigma_1, \ldots, \sigma_n)$. Assuming that the sample is randomly sorted as σ_p the kth dimensional feature $x_{\sigma_p,k}$ is treated as a categorical feature.

$$\hat{x}_k^i = \frac{\sum_{j=1}^{p-1} \left[x_{\sigma_j,k} = x_{\sigma_p,k}\right] . Y_{\sigma_j} + a.p}{\sum_{j=1}^{p-1} \left[x_{\sigma_j,k} = x_{\sigma_p,k}\right] + a} \quad (7)$$

Here, $\hat{x}_k^i$ represents a target variable statistic, $x_{\sigma_j,k}$ denotes a categorical feature, p signifies the a priori value, Y_{σ_j} represents the label value of the respective feature, and $a > 0$ serves as a weight denoting the priori value [20].

4.2.4 Multilayer perceptron

Multilayer Perceptron (MLP) is a variant of artificial neural network architecture [21]. It is structured with an input layer, hidden layers, and an output layer. Nodes or neurons are present in each layer. The weights between neurons represent the strength of connections, while biases enable the model to consider variations. Weights and biases are adjusted using the optimization algorithms to reduce the loss function. Eq. (8) and Eq. (9) describe a neuron k.

$$u_k = \sum_{j=1}^{n} w_{kj} x_j \quad (8)$$

$$y_k = \varphi(u_k + b_k) \quad (9)$$

where $x_1, x_2, \ldots, x_n$ denote the input signals and $w_{k1}, w_{k2}, \ldots, w_{kn}$ denote the weights of neuron k; b_k is the bias, φ is the activation function, and y_k is the output signal of neuron k [22].

4.2.5 Implementation

Jupyter Notebook serves as the Integrated Development Environment for all operations using Python (version 3.11.3). Python libraries including Seaborn, Scikit-Learn, Pandas, Matplotlib, and NumPy were employed. System specifications utilized in ML algorithm generation are presented in Table 1.

Table 1. System specifications

Hardware configuration for ML algorithms generation	
RAM	8GB
CPU	12th Gen Intel(R) Core (TM)i5-1235U, 1.30 GHz, 10 Core(s), 12 Logical Processor(s)
Storage	512GB SSD
GPU	Integrated GPU

V. RESULTS AND DISCUSSIONS

5.1 Train and test dataset

All classifiers are initially trained on the training dataset, and subsequently, the classifiers performance analyzed using the test dataset. The dataset is split into an 80:20 ratio, allocating 80% for training and the remaining 20% for testing, consisting of 424 instances in the training set and 107 instances in the testing set. In Fig. 6, the confusion matrix is presented, depicting the association between the actual class and the predicted class.

5.2 Accuracy

It calculates the proportion of accurately predicted instances relative to the total number of instances in a dataset. Eq. (10) gives the accuracy scores for individual classifiers. Fig. 7 illustrates the comparison of accuracy scores of base classifiers Vs. proposed WMVE classifier.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{10}$$

5.3 Precision

It assesses the model's capability to accurately identify true positives within all instances it predicted as positive. Eq. (11) gives the precision of the model.

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

5.4 Recall

It quantifies the proportion of actual positive classes correctly predicted by the model. Recall of the model is described in Eq. (12).

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

5.5 F1 Score

It is expressed as the harmonic average of precision and recall of the model with a desirable value of 1 and an undesirable value of 0. Eq. (13) provides the F1 score of the model.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{13}$$

5.6 ROC curve

An ROC curve known as receiver operating characteristic curve [23], is a pictorial representation that depicts the classification algorithm performance. It illustrates the balance between a model's true positive rate (sensitivity) and its false positive rate (1-specificity) at different probability thresholds. Through the execution of soft voting, probability estimates for each class are provided by each individual classifier in the ensemble. This feature is valuable for calculating the ROC curve, as it allows a detailed examination of the classifier performance across different probability thresholds.

The AUROC (Area Under the Receiver Operating Characteristic Curve) was calculated using the trapezoidal rule as described in Eq. (14). Eq. (15) and Eq. (16) give the true positive rate (TPR) and false positive rate (FPR) values. Fig. 9 illustrates the ROC plot of individual classifiers with their corresponding AUROC values.

$$AUROC = \sum_{i=1}^{N} TPR(i) * \Delta FPR(i) \tag{14}$$

$$TPR(Recall) = \frac{TP}{TP + FN} \tag{15}$$

$$FPR = \frac{FP}{FP + TN} \tag{16}$$

$$\Delta: sampling\ interval$$

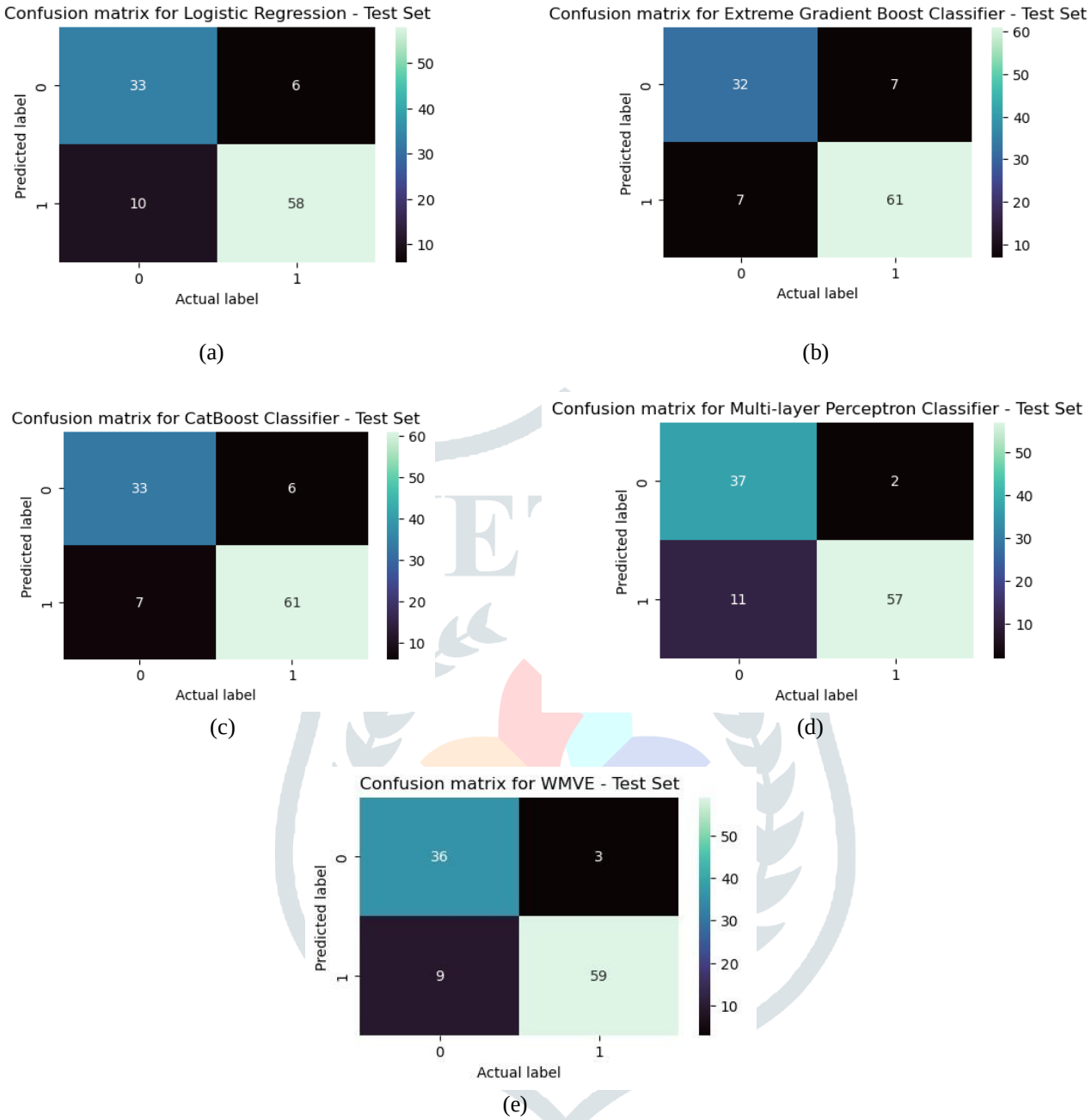

Fig. 6. Confusion matrix of (a) LR (b) XGBoost (c) CatBoost (d) MLP (e) WMVE.

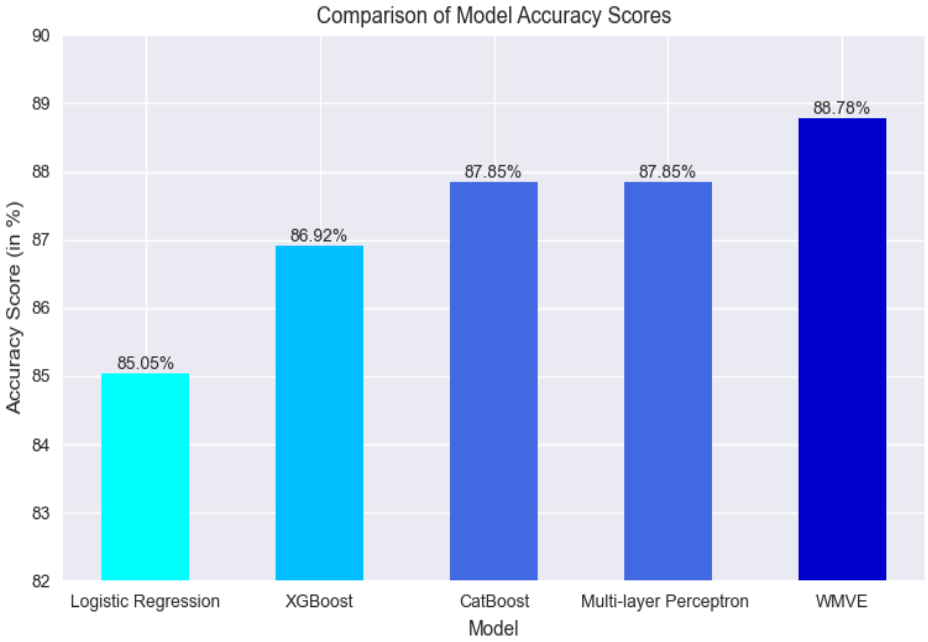

Fig. 7. Comparison of accuracy scores: base classifier Vs. proposed WMVE classifier.

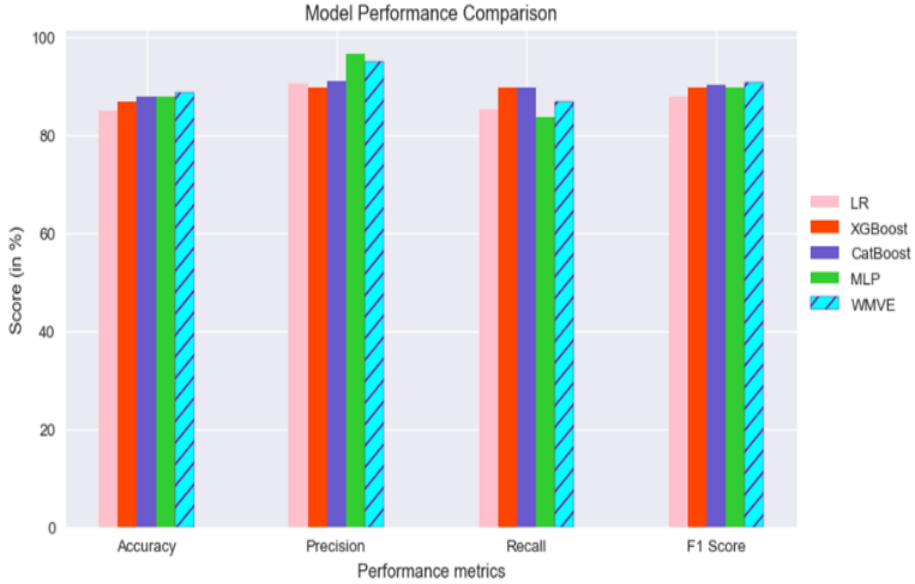

Fig. 8. Performance evaluation: base classifier Vs. proposed WMVE classifier.

A confusion matrix for the WMVE classifier in Fig. 6 (e) revealed the accurate identification of 36 cases with CVD (true positives) and the precise classification of 59 instances without CVD (true negatives). However, the WMVE model displayed limitations, misclassifying 3 cases as false positives and overlooking 9 cases with CVD (false negatives). In the comparative analysis between the base classifiers and the proposed WMVE classifier in Fig. 7, the WMVE classifier demonstrated an improvement in accuracy, achieving a score of 88.78%. This comparison highlights the efficacy of the WMVE classifier in enhancing predictive accuracy compared to its constituent base classifiers.

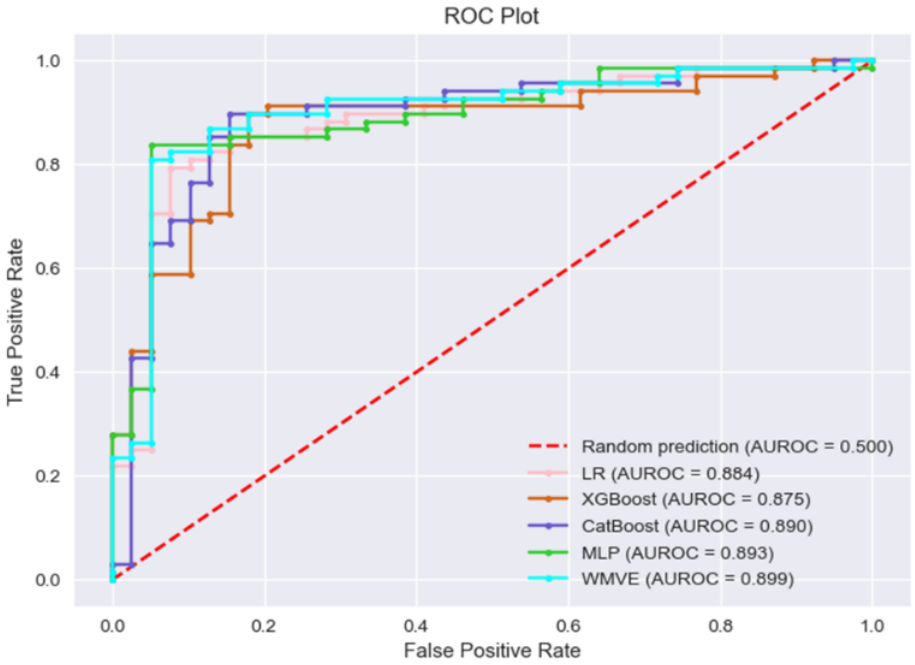

Fig. 9. ROC curves.

A comprehensive performance evaluation between the base classifiers and the proposed WMVE classifier is presented in Fig. 8. Regarding precision, the MLP exhibits the highest performance with a score of 96.61%. For recall, XGBoost and CatBoost both achieve a score of 89.70%. Finally, when evaluating the F1 score, the WMVE classifier outperforms others with a score of 90.76%. Table 2 represents the accuracy comparison among existing cardiovascular disease (CVD) prediction models. In the proposed model, the assignment of weights to the base classifiers lacks precise adaptive adjustment based on the individual performance of each classifier, and this limitation restricts the potential for increased accuracy. To address this, adaptive weight adjustments could be incorporated based on the performance of the individual base classifiers.

Table 2. Comparative accuracy analysis of existing CVD risk prediction classifiers

Author	ML Algorithm	Accuracy (%)
B. A. Tama et al. [7]	Two-tier ensemble with feature selection using PSO	85.71
Reddy et al. [8]	Combination of Chi-squared and SMO	86.46
C. Gazeloğlu [9]	Feature selection using correlation with NB	84.81
C.B.C Latha et al. [10]	Hard voting involving MLP, BN, NB, and RF	85.48
Proposed method	**Weighted Majority Voting Ensemble (WMVE)**	**88.78**

VI. CONCLUSION

This study has demonstrated the implementation of a WMVE approach for improved accuracy in CVD prediction. The WMVE, consisting of four ML models, achieved an accuracy of 88.78% and an AUROC of 0.899, with a higher weight assigned to the Multilayer perceptron classifier. The Pearson correlation coefficient was employed to identify the highly correlated features. This work can be extended to explore adaptive weight adjustments in the WMVE classifier and have a comparative study with additional ensemble techniques, which could lead to enhanced accuracy in the diagnosis of heart disease. Our approach empowers healthcare professionals to promptly make critical decisions about patient well-being.

REFERENCES

[1] "Cardiovascular diseases (CVDs)." WHO. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed May. 17, 2024).
[2] "About Heart Disease." CDC. https://www.cdc.gov/heartdisease/about.htm (accessed May. 17, 2024).
[3] "Cardiovascular disease." NHS. https://www.nhs.uk/conditions/cardiovascular-disease/ (accessed May. 17, 2024).
[4] D. Stojanov, E. Lazarova, E. Veljkova, P. Rubartelli, and M. Giacomini, "Predicting the outcome of heart failure against chronic-ischemic heart disease in elderly population – Machine learning approach based on logistic regression, case to Villa Scassi hospital Genoa, Italy," Journal of King Saud University – Science, vol. 35, no. 3, Apr. 2023, doi: 10.1016/j.jksus.2023.102573.
[5] I. Tougui, A. Jilbab, and J. El Mhamdi, "Heart disease classification using data mining tools and machine learning techniques," Health and Technology (Berl), vol. 10, pp. 1137–1144, 2020, doi: 10.1007/s12553-020-00438-1.

[6] A. Ghasemieh, A. Lloyed, P. Bahrami, P. Vajar, and R. Kashef, "A novel machine learning model with Stacking Ensemble Learner for predicting emergency readmission of heart-disease patients," Decision Analytics Journal, vol. 7, Jun. 2023, doi: 10.1016/j.dajour.2023.100242.

[7] B. A. Tama, S. Im, and S. Lee, "Improving an Intelligent Detection System for Coronary Heart Disease Using a Two-Tier Classifier Ensemble," BioMed Research International, vol. 2020, 2020, doi: 10.1155/2020/9816142.

[8] K. V. V. Reddy, I. Elamvazuthi, A. A. Aziz, S. Paramasivam, H. N. Chua, and S. Pranavanand, "Heart disease risk prediction using machine learning classifiers with attribute evaluators," Applied Sciences (Switzerland), vol. 11, no. 18, Sep. 2021, doi: 10.3390/app11188352.

[9] C. Gazeloğlu, "Prediction of heart disease by classifying with feature selection and machine learning methods," Progress in Nutrition, vol. 22, no. 2, pp. 660–670, 2020, doi: 10.23751/pn.v22i2.9830.

[10] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," Informatics in Medicine Unlocked, vol. 16, Jan. 2019, doi: 10.1016/j.imu.2019.100203.

[11] D. Ananey-Obiri and E. Sarku, "Predicting the Presence of Heart Diseases using Comparative Data Mining and Machine Learning Algorithms," International Journal of Computer Applications, vol. 176, April 2020, doi: 10.5120/ijca2020920034.

[12] K. Kanagarathinam, D. Sankaran, and R. Manikandan, "Machine learning-based risk prediction model for cardiovascular disease using a hybrid dataset," Data & Knowledge Engineering, vol. 140, Jul. 2022, doi: 10.1016/j.datak.2022.102042.

[13] Heart Disease, UCI Machine Learning Repository [Dataset] [Online]. Available: https://doi.org/10.24432/C52P4X.

[14] Z. Li, Y. Yang, L. Li, and D. Wang, "A weighted Pearson correlation coefficient based multi-fault comprehensive diagnosis for battery circuits," Journal of Energy Storage, vol. 60, Apr. 2023, doi: 10.1016/j.est.2022.106584.

[15] F. Ye, Z. Zhang, K. Chakrabarty, and X. Gu, "Board-level functional fault diagnosis using artificial neural networks, support-vector machines, and weighted-majority voting," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 32, no. 5, pp. 723–736, 2013, doi: 10.1109/TCAD.2012.2234827.

[16] M. Fang and D. T. N. Huy, "Building a cross-border e-commerce talent training platform based on logistic regression model," Journal of High Technology Management Research, vol. 34, no. 2, Nov. 2023, doi: 10.1016/j.hitech.2023.100473.

[17]"XGBoost." nvidia. Available: https://www.nvidia.com/en-us/glossary/data-science/xgboost/#:~:text=XGBoost%20is%20a%20scalable%20and,model%20performance%20and%20computational%20speed. (accessed May. 25, 2024).

[18] Z. H. Wang, Y. F. Liu, T. Wang, J. G. Wang, Y. M. Liu, and Q. X. Huang, "Intelligent prediction model of mechanical properties of ultrathin niobium strips based on XGBoost ensemble learning algorithm," Computational Materials Science, vol. 231, Jan. 2024, doi: 10.1016/j.commatsci.2023.112579.

[19] "Preliminary calculation of splits." CatBoost. https://catboost.ai/en/docs/concepts/algorithm-main-stages_pre-count (accessed May. 25, 2024).

[20] S. Zhang, X. Lu, and Z. Lu, "Improved CNN-based CatBoost model for license plate remote sensing image classification," Signal Processing, vol. 213, Dec. 2023, doi: 10.1016/j.sigpro.2023.109196.

[21] "Multi-layer Perceptron." scikit-learn. https://scikit-learn.org/stable/modules/neural_networks_supervised.html (accessed May. 25, 2024).

[22] M. Alviano, F. Bartoli, M. Botta, R. Esposito, L. Giordano, and D. Theseider Dupré, "A preferential interpretation of MultiLayer Perceptrons in a conditional logic with typicality," International Journal of Approximate Reasoning, vol. 164, Jan. 2024, doi: 10.1016/j.ijar.2023.109065.

[23] A. E. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," Pattern Recognition, vol. 30, no. 7, pp. 1145-1159, 1997, doi: 10.1016/S0031-3203(96)00142-2.