Optimizing Cardiovascular Disease Detection Using Ranking-Based Feature Selection Machine Learning Models

Anuradha S. Deokar
Computer Engineering Department, AISSMSCOE, Kennedy Road, SPPU University, Pune, India
asdeokar@aissmscoe.com (corresponding author)

Madhavi A. Pradhan
Computer Engineering Department, AISSMSCOE, Kennedy Road, SPPU University, Pune, India
mapradhan@aissmscoe.com

Received: 5 May 2025 | Revised: 28 May 2025, 25 June 2025, 10 July 2025, and 6 August 2025 | Accepted: 22 August 2025

 DOI: https://doi.org/10.48084/etasr.11923

ABSTRACT

Cardiovascular Disease (CVD) continues to be one of the leading causes of mortality worldwide, emphasizing the urgent need for accurate and efficient diagnostic solutions. This study presents a machine learning-based framework for the classification of CVD that is transparent, reliable, and computationally optimized. The performance of Machine Learning (ML) models can be significantly hindered by class imbalances and high-dimensional datasets. To address these challenges, various Feature Selection (FS) techniques were applied to identify the most informative predictors, thereby reducing both dimensionality and computational complexity. A comprehensive evaluation of multiple ML algorithms was conducted in conjunction with FS methods, using standard performance metrics such as accuracy, precision, recall, and F1-score. The results demonstrate that integrating effective feature selection techniques not only improves model interpretability but also mitigates the risk of overfitting. The proposed system incorporates a novel feature ranking algorithm that enhances the selection of optimal predictors, playing a crucial role in the construction of robust and reliable diagnostic models. The proposed ensemble model achieved a peak accuracy of 98.20% with an 80:10:10 split for training, testing, and validation. These findings highlight the potential of the proposed model to support early clinical decision-making, enabling timely interventions and reducing the likelihood of severe CVD-related outcomes. The integration of such intelligent systems into clinical workflows may contribute significantly to improving patient care and disease management.

Keywords-cardiovascular disease; machine learning; feature selection; ranking algorithm; interpretability

I. INTRODUCTION

Cardiovascular Disease (CVD) is one of the leading causes of global mortality, with the WHO projecting more than 23.6 million deaths annually by 2030 [1]. Many CVD cases are due to changeable risk factors, such as diabetes, high blood pressure, and sedentary lifestyles. Early detection is critical and can significantly improve outcomes. In recent years, ML has become a pivotal tool in advancing CVD prediction by improving diagnostic accuracy and supporting clinical decision-making. Among the various approaches, ensemble learning and feature selection techniques have emerged as particularly effective in optimizing model performance and interpretability. Advanced ensemble models combining feature selection methods such as SHAP-RFE (SSHO) and gradient-sharing mechanisms (DGSM) between CatBoost and neural networks have demonstrated superior predictive accuracy, making them effective tools in CVD risk prediction [2]. Recent research highlights the importance of handling class imbalance in clinical data using resampling techniques to enhance the performance of ML models [3].

A notable ensemble-based framework utilized chi-square feature selection to refine the Cleveland heart disease dataset, successfully identifying five critical features from 13 clinical attributes. By aggregating the predictions of classifiers such as Naïve Bayes (NB), Random Forest (RF), and Linear Regression (LR), a precision of 92.11% was achieved, while simultaneously reducing computational overhead by more than 50%, thus demonstrating both efficiency and accuracy in CVD risk prediction [4]. Further advances have been made using ensemble stacking methods, integrating diverse base models including RF, Decision Trees (DT), XGBoost, Gaussian NB, and LightGBM. These models, trained on a balanced dataset enhanced using the SMOTE technique and normalized with z-score, yielded a high precision rate of 98.4%, underscoring the efficacy of ensemble stacking in managing class imbalance and boosting predictive reliability [5].

In [6], a stacking ensemble classifier was developed using the Johnson transformation for data normalization and a feature selection process. This model combined three base-level classifiers through an aggregation layer utilizing the Dependent Ordered Weighted Averaging (DOWA) operator. Comparative evaluations revealed that models such as SVM and RF outperformed traditional classifiers, emphasizing the value of methods that continue to improve CVD diagnosis by integrating intelligent feature selection strategies. In [7], DT, SVM, and boosting algorithms were applied, demonstrating that these methods, particularly when paired with relevant feature selection, significantly improve diagnostic precision. In [8], a self-supervised learning framework fused cardiac MRI (CMRI), ECG signals, and clinical records. This approach showcased the potential of unsupervised feature learning in improving classification outcomes and emphasized the transformative role of DL. Principal Component Analysis (PCA) has also been used effectively to improve model robustness. In [9], a machine learning framework leveraging PCA for early coronary heart disease detection achieved an accuracy of 96.12% across multiple classifiers, highlighting its utility in enhancing generalization and reducing overfitting in complex clinical datasets.

This work builds on these advances by incorporating several key innovations: (i) addressing class imbalance through data augmentation techniques, (ii) applying a novel feature ranking algorithm to identify and prioritize critical clinical variables, and (iii) validating the method on the Framingham Heart Study dataset [10]. Experimental results demonstrate the superiority of the proposed system over conventional ML models, offering a reliable and scalable approach to the early prediction of CVD risk.

II. PROPOSED METHOD

The proposed system is intended to evaluate the performance of CVD prediction, determining whether a patient has CVD and identifying CVD risk factors. Figure 1 shows details of the system architecture.

A. Preprocessing

The Framingham dataset [10] has 15 attributes, 1 target attribute, and 4240 instances, containing numerical and categorical features, along with missing values. The empty numerical values were filled with the median of the column, and the categorical columns with the mode.

Data preprocessing is a critical step in an ML pipeline to ensure the quality and reliability of the input data. It involves addressing missing values, eliminating outliers, and standardizing the dataset to improve model performance. In scenarios where class imbalance exists—commonly observed in medical datasets—Synthetic Minority Over-Sampling Technique (SMOTE) is employed to balance the class distribution. By generating synthetic samples for the minority class, SMOTE enhances the model's ability to learn from underrepresented instances, thereby improving overall predictive accuracy and robustness.

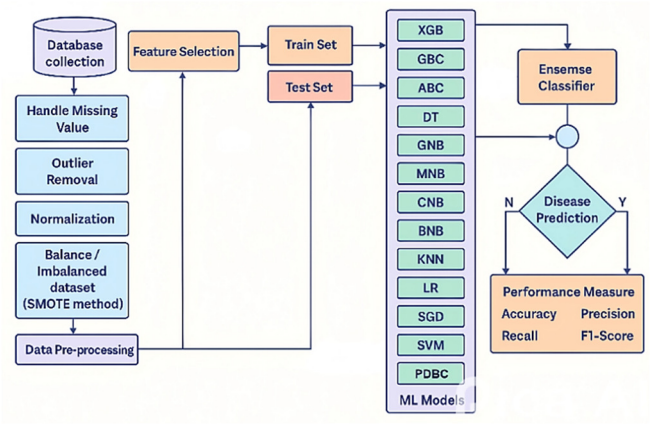

Fig. 1. Proposed architecture model

Normalization divides each data value by the maximum value to scale it. This method transforms the data into a range between 0 and 1, preserving the relative distribution while minimizing the impact of varying scales, and is commonly used to enhance the performance and convergence of ML models.

$$data_{normalized} = \frac{data}{\max(data)} \quad (1)$$

B. Training Dataset

Once the data has been preprocessed, it must be partitioned into training and testing sets. It is essential to ensure that the training dataset encompasses a diverse and representative range of feature values to support robust model learning. To achieve this, cross-validation techniques are employed, enabling the model to generalize well by minimizing bias and variance during the training process [11]. The dataset is split into training, validation, and testing datasets in an 80:10:10 ratio.

C. Algorithm Selection and Evaluation

Predictive accuracy is used to evaluate the algorithms. True positives, false positives, true negatives, and false negatives are given by a confusion matrix. Accuracy, F1-score, Error rate, and Recall served as performance evaluation metrics [11].

This study used an ensemble classifier. A gradient boosting approach was applied to avoid bias (e.g., underfitting) and improve its predictive performance.

III. FEATURE SELECTION

Feature selection helps in building faster, more accurate, and interpretable models [12]. It prevents overfitting, reduces complexity**,** and improves efficiency, while retaining original feature meanings. Different methods (Filter, Wrapper, Embedded) work for different types of data and models.

A. Filter Methods (Statistical Techniques)

In these methods, features are selected for their association with the variable of interest.

1) Mutual Information (for both Categorical and Continuous)

Mutual information (MI) measures the dependence between variables [13]. It's a non-parametric method, meaning that it can capture non-linear relationships. MI (2) determines how

much information is acquired about one stochastic variable by observing another. If two variables are not dependent, their MI is zero. The greater the mutual information, the more dependent the variables are.

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \quad (2)$$

where $H(X)$ is the entropy of X and $H(X|Y)$ is the conditional entropy of X given Y. Figure 2 shows the MI scores for each feature, sorted in descending order of importance. The x-axis shows the MI scores for each feature. Higher scores (age, hypertension, sysBP, diabetes, etc.) indicate a stronger relationship between the feature and the TenYearCHD target variable. The y-axis shows the feature name.

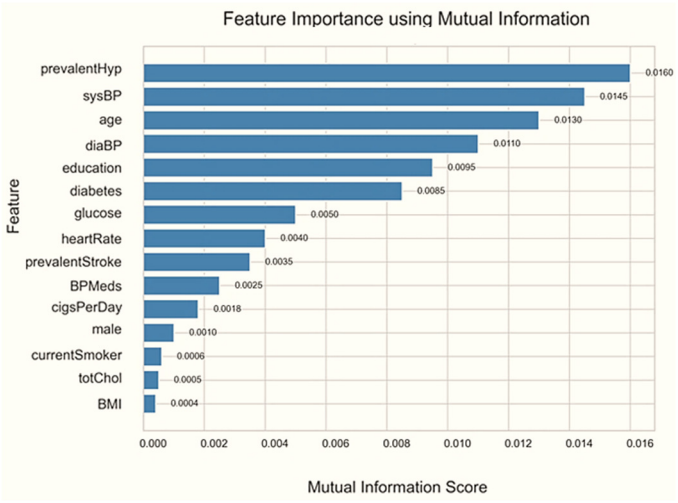

Fig. 2. Mutual information graph.

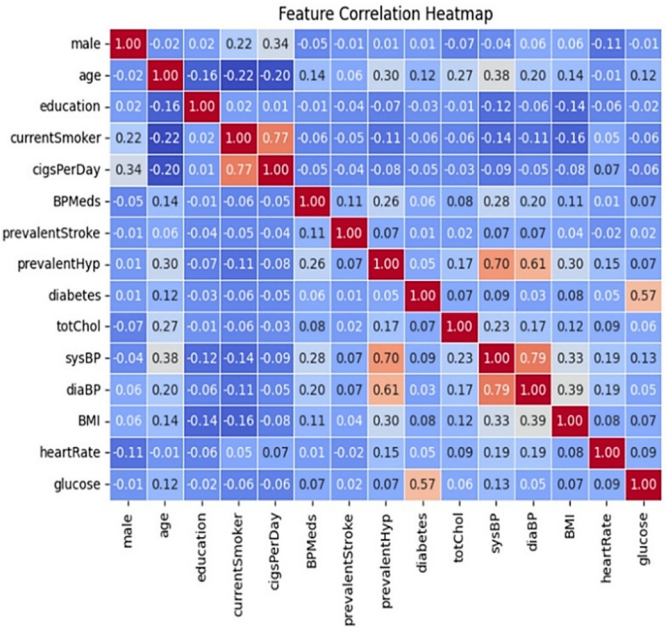

Fig. 3. Feature correlation graph.

2) Feature Correlation Analysis

Correlation analysis evaluates the linear relationship between features, as shown in Figure 3, is calculated using:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \quad (3)$$

where μ is the population dataset mean, N is the sum of observations, x_i denotes individual data points, and σ^2 is the population variance. A favorable association indicates that when one variable increases, the other usually does as well. When one variable tends to increase while the other tends to decrease, this is known as a negative correlation.

B. Embedded Methods

Embedded methods perform feature selection as part of the model training process [14-16].

1) LASSO Regression

LASSO (Least Absolute Shrinkage and Selection Operator) is a linear model that performs both variable selection and regularization to enhance prediction accuracy and interpretability. LASSO assigns a coefficient to each feature in the dataset. Crucially, it shrinks some of these coefficients to exactly zero. By observing which features have non-zero coefficients, a list of the highest significant predictors can be created. This list represents the part of the features that LASSO has identified as relevant for predicting CVD risk.

LASSO uses a regularization parameter called lambda (often called alpha in scikit-learn). The value of this parameter influences the output. A higher lambda value results in more coefficients being shrunk to zero, leading to a simpler model. The output may consist of the chosen lambda value and in what way it was selected (e.g., using cross-validation).

The LASSO regression objective function is:

$$min_{\beta_0,\beta} \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p}|\beta_j| \quad (4)$$

where y_i is the target variable for the i^{th} observation, x_{ij} is the feature j for the i^{th} observation, β_0 is an intercept term, β_j denotes the coefficients of the feature, n is the number of observations (samples), p is the number of features (predictors), and λ is a regularization parameter controlling shrinkage, where:

$$if \begin{cases} \lambda = 0, & \text{it behaves like ordinary least squares.} \\ \lambda \text{ is large}, & \text{more coefficients shrink to zero.} \end{cases} \quad (5)$$

In essence, the key outputs are which features are deemed important (those with non-zero coefficients). Figure 4, shows important features such as age, sysBP, male, and cigsPerDay for the risk factor of CVD.

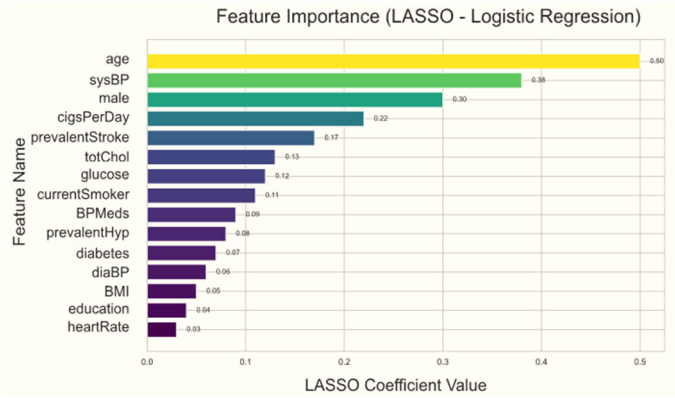

Fig. 4. LASSO graph.

2) Gradient Boosting Model (GBM)

The general formulation of a GBM is:

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

where $F_m(x)$ is the prediction at iteration m, $F_{m-1}(x)$ is the prediction from the previous step, $h_m(x)$ is a weak learner (usually a decision tree) fitted to the negative gradient of the loss function, γ is the learning rate (step size) that controls how much to update the model, and m is the number of boosting iterations.

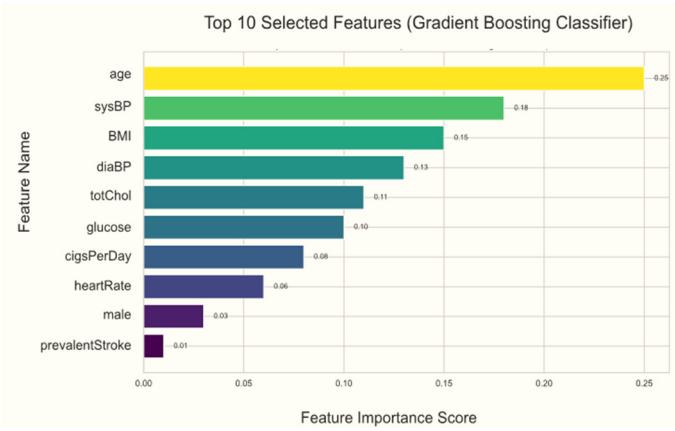

Fig. 5. GBM graph.

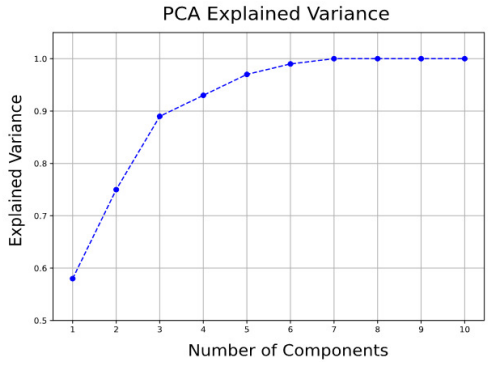

Fig. 6. PCA graph.

C. Dimensionality Reduction

1) Principal Component Analysis (PCA)

PCA reduces the dimensionality of the data while retaining as much variance as possible [15], as shown in Figure 6.

IV. MODEL INTERPRETABILITY

SHAP values provide a unified approach to interpreting the output of ML models by assigning each feature a contribution score for a given prediction [14]. In CVD prediction, SHAP values help in understanding how individual risk factors—such as age, blood pressure, cholesterol, and smoking—impact the model's decisions for each patient. This interpretability increases trust in the model and helps clinicians to identify the most influential features contributing to a diagnosis, supporting informed medical decision-making.

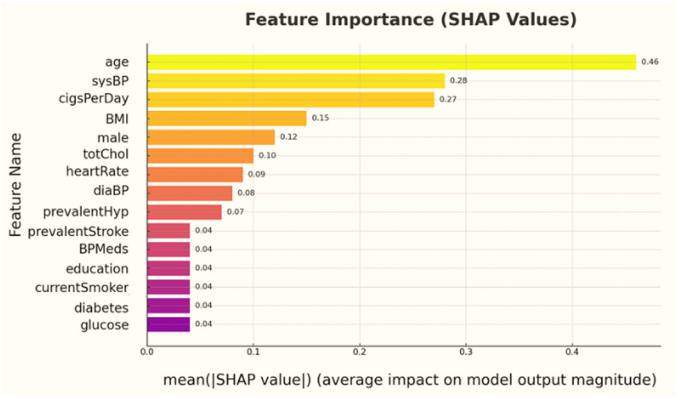

Fig. 7. Graph of SHAP values.

V. PROPOSED ALGORITHMS AND MODEL OPTIMIZATION

This algorithm prioritizes feature selection models for CVD prediction, ensuring high predictive accuracy while addressing the specific challenges of medical datasets.

```
Algorithm 1: Proposed ranking algorithm
for CVD Prediction
STEP 1: Input dataset
  - A dataset containing n features and a
    target class label.
  - The number of top-ranked features to
    select (optional).
STEP 2: Preprocess the data:
   - Clean and normalize data.
   - Encode categorical variables if
     necessary.
STEP 3: Compute a score:
   - Compute a score for each feature
     that reflects its importance or
     correlation with the target
     variable.
   - Feature selection metrics include:
        o Information Gain
        o Chi-Square statistics
        o Mutual Information
```

```
        o  Pearson Correlation (for
           numeric data)
STEP 4: Rank features:
   -  Sort the features based on their
      computed scores in descending order.
STEP 5: Select features:
   -  Select the top k features based on
      the sorted list.

Algorithm 2: Updating decision boundary
INPUT: Heart Disease Dataset
Step 1: Exploratory Data Analysis (EDA):
   -  For each type of dataset (balanced
      and unbalanced):
   -  Examine data distribution,
      structure, and attributes.
   -  Identify potential correlations and
      trends.
   -  Compare feature distributions across
      different categories.
STEP 2: Preprocess the data:
   -  Apply imputation techniques or other
      necessary methods.
   -  Perform feature selection or assess
      feature importance.
STEP 3: Build and Train Models:
   -  Train models.
   -  Validate models on the testing set.
   -  Evaluate model performance using
      proper evaluation metrics.
   -  Identify models with the best
      performance.
STEP 4: Final model evaluation:
   -  Generate final predictions using the
      selected models.
   -  Evaluate the final classification
      results using the following metrics:
```

- $Accuracy = \frac{TP+TN}{FP+FN+TP+TN}$
- $Precision = \frac{TP}{FP+TP}$
- $Recall = \frac{TP}{FN+TP}$
- $F1-Score = \frac{2\times(Precision\times Recall)}{Precision+Recall}$

where TP denotes True Positives, TN denotes True Negatives, FP denotes False Positives, and FN denotes False Negatives.

VI. RESULTS

The proposed ensemble model incorporates classifiers like XGBoost and SVM. An updating mechanism for the decision boundary is proposed that dynamically adjusts classification thresholds according to misclassified instances for enhancing predictive accuracy. Feature selection improves accuracy and decreases training time. Table I shows the performance results of different models tested with various feature selection techniques.

A Receiver Operating Characteristic (ROC) curve is a graph that shows the relationship between a model's True Positive Rate (TPR) and False Positive Rate (FPR) at different thresholds. The TPR is plotted on the y-axis, and the FPR is plotted on the x-axis. The Area Under the ROC curve (AUC-ROC) evaluates model performance by plotting the true positive rate against the false positive rate and computing the area beneath the curve. A score closer to 1 indicates a stronger ability of the algorithm to differentiate between the two outcome classes. The ROC AUC score ranges from 0 to 1, where 0.5 indicates random guessing, and 1 indicates perfect performance. Figure 8 shows the ROC curve for the proposed model.

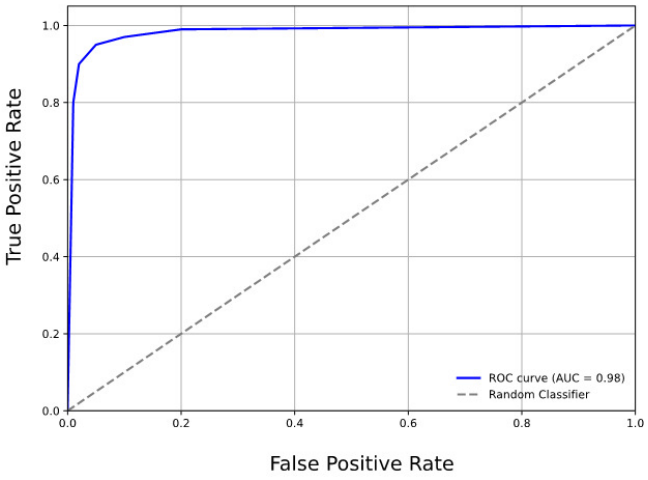

Fig. 8. ROC curve with FPR and TPR.

The training time before and following feature selection decreased, from 5.3 s to 4 s for the balanced dataset, also improving accuracy from 96.42% to 98.65%. Table II shows the training time comparison for the balanced and imbalanced datasets, with and without applying feature selection techniques. Table III shows the comparative analysis of the results for the proposed model with those of previous studies.

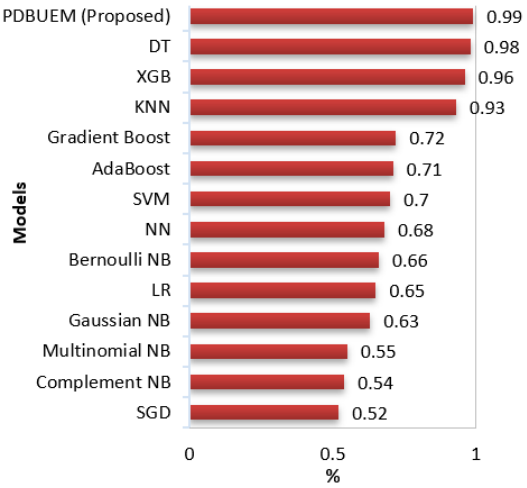

Fig. 9. Accuracy comparison with traditional ML methods.

TABLE I. COMPARISON OF VARIOUS MODELS WITH FS METHODS, FEATURES, AND ACCURACY

	Model	Algorithm/ classifier	Feature selection method	No. of features	Selected features	Accuracy	Precision	Recall	F1-score
1	Tree-based Models	Gradient Boost	Feature Importance	15	10	71.78	71.81	71.78	71.77
		AdaBoost	Feature Importance	15	8	70.89	70.90	70.89	70.88
		XGBoost	Feature Importance	15	10	97.25	97.27	97.25	97.25
		Decision Tree	SHAP Values	15	10	96.80	96.85	96.80	96.80
		PDBUEM	SHAP Values	15	10	98.65	98.25	98.20	98.20
2	Linear Models	SVM	Recursive Feature Elimination (RFE)	15	9	67.50	67.60	67.50	67.45
		LR	L1 Regularization (LASSO)	15	10	66.90	66.90	66.90	66.90
		SGD	Mutual Information	15	10	62.45	62.38	66.45	66.65
3	Bayesian models (Naïve Bayes)	Gaussian NB	Mutual Information	15	10	63.30	63.60	63.30	63.20
		MNB	Chi-Square	15	10	54.70	54.60	54.65	54.64
		CNB	Chi-Square	15	10	54.60	54.55	54.58	54.65
		BNB	Variance Threshold,	15	10	64.40	64.55	64.35	64.25
4	Distance-Based Models	KNN	PCA	15	10	91.70	92.30	91.68	91.64
5	Neural Networks	NN	SHAP	15	10	67.50	67.70	67.48	67.40

TABLE II. TRAINING TIME BEFORE AND AFTER FEATURE SELECTION FOR BALANCED AND IMBALANCED DATASETS

Parameters	Before selecting features	After feature selection
Imbalanced dataset		
Training time	4.7 s	3.8 s
No of features	15	10
Accuracy	92.95	94.21%
Balanced dataset		
Training time	5.3 s	4 s
No of features	15	10
Accuracy	96.42%	98.65%

TABLE III. COMPARATIVE ANALYSIS WITH SIMILAR METHODS

Ref.	Feature selection method	Best model	Accuracy
[3]	No explicit selection.	RF	94.31%
[4]	Chi-square	Voting ensemble (NB, LR, RF, KNN)	92.11%
[7]	RFE	DT+NB	85.63%
Proposed (PDBUEM)	Feature Importance	Ensemble classifier	98.40%

VII. CONCLUSION AND FUTURE WORK

The main objective of this study was to use ranking-based feature selection and ML models to predict CVD risk. Different ML algorithms often require tailored feature selection approaches to achieve optimal performance. The proposed ranking algorithm systematically evaluates and prioritizes features, enabling the selection of the most significant predictors while reducing computational complexity. By emphasizing clinically relevant attributes, the model improves interpretability, trust, and transparency, which are key factors in medical diagnostics. The integration of this method led to a notable improvement in model performance, with accuracy reaching 98.65%. In addition to high accuracy, the system demonstrates excellent efficiency in terms of prognosis speed. Evaluation metrics such as precision, recall, and F1-score further validate the robustness of the model, consistently achieving values around 98.25% and 98.20%.

Based on computational evaluations, the proposed approach is suitable for both large-scale implementations and real-time applications. In the future, developing models that adapt to individual patient profiles, considering age, ethnicity, and comorbidities, can lead to more personalized and clinically relevant predictions.

REFERENCES

[1] World Health Organization, "Global status report on noncommunicable diseases 2010," World Health Organization, Geneva, Switzerland, 2011. [Online]. Available: https://iris.who.int/handle/10665/44579.

[2] O. C. Gold and J. Lawrence, "Ensemble of CatBoost and neural networks with hybrid feature selection for enhanced heart disease prediction," *The Scientific Temper*, vol. 15, no. 04, pp. 3157–3164, Dec. 2024, https://doi.org/10.58414/SCIENTIFICTEMPER.2024.15.4.24.

[3] M. Rahardi, A. Aminuddin, F. F. Abdulloh, B. P. Asaddulloh, H. R. Enriquez, and K. Kusnawi, "Analyzing the Impact of Data Resampling on Stroke Prediction using Machine Learning," *Engineering, Technology & Applied Science Research*, vol. 15, no. 2, pp. 20790–20797, Apr. 2025, https://doi.org/10.48084/etasr.9736.

[4] A. E. Korial, I. I. Gorial, and A. J. Humaidi, "An Improved Ensemble-Based Cardiovascular Disease Detection System with Chi-Square Feature Selection," *Computers*, vol. 13, no. 6, May 2024, Art. no. 126, https://doi.org/10.3390/computers13060126.

[5] M. Aruna and V. B. Shalini, "FO-RS-3TM: A High-Performance Ensemble Model for Heart Disease Prediction with Feature Selection and Hyperparameter Tuning," in *2024 8th International Conference on Inventive Systems and Control (ICISC)*, Coimbatore, India, Jul. 2024, pp. 99–105, https://doi.org/10.1109/ICISC62624.2024.00024.

[6] M. H. Chagahi, S. M. Dashtaki, B. Moshiri, and M. D. J. Piran, "Cardiovascular disease detection using a novel stack-based ensemble classifier with aggregation layer, DOWA operator, and feature transformation," *Computers in Biology and Medicine*, vol. 173, May 2024, Art. no. 108345, https://doi.org/10.1016/j.compbiomed.2024.108345.

[7] S. M. Alanazi and G. S. M. Khamis, "Optimizing Machine Learning Classifiers for Enhanced Cardiovascular Disease Prediction," *Engineering, Technology & Applied Science Research*, vol. 14, no. 1, pp. 12911–12917, Feb. 2024, https://doi.org/10.48084/etasr.6684.

[8] F. Girlanda, O. Demler, B. Menze, and N. Davoudi, "Enhancing Cardiovascular Disease Prediction through Multi-Modal Self-Supervised Learning." arXiv, 2024, https://doi.org/10.48550/ARXIV.2411.05900.

[9] M. J. J. Ghrabat *et al.*, "Utilizing Machine Learning for the Early Detection of Coronary Heart Disease," *Engineering, Technology &*

Applied Science Research, vol. 14, no. 5, pp. 17363–17375, Oct. 2024, https://doi.org/10.48084/etasr.8171.

[10] "Framingham heart study dataset." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset.

[11] S. Pattanayak and T. Singh, "Cardiovascular Disease Classification Based on Machine Learning Algorithms Using GridSearchCV, Cross Validation and Stacked Ensemble Methods," in *Advances in Computing and Data Sciences*, 2022, pp. 219–230, https://doi.org/10.1007/978-3-031-12638-3_19.

[12] P. Singh, G. K. Pal, and S. Gangwar, "Prediction of Cardiovascular Disease Using Feature Selection Techniques," *International Journal of Computer Theory and Engineering*, vol. 14, no. 3, pp. 97–103, 2022, https://doi.org/10.7763/IJCTE.2022.V14.1316.

[13] S. A. Sabab, Md. A. R. Munshi, A. I. Pritom, and Shihabuzzaman, "Cardiovascular disease prognosis using effective classification and feature selection technique," in *2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec)*, Dhaka, Bangladesh, Dec. 2016, pp. 1–6, https://doi.org/10.1109/MEDITEC.2016.7835374.

[14] S. M. Ganie, P. K. D. Pramanik, and Z. Zhao, "Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets," *Scientific Reports*, vol. 15, no. 1, Apr. 2025, Art. no. 13912, https://doi.org/10.1038/s41598-025-97547-6.

[15] G. S. M. Khamis and S. M. Alanazi, "Exploring sex disparities in cardiovascular disease risk factors using principal component analysis and latent class analysis techniques," *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, May 2023, Art. no. 101, https://doi.org/10.1186/s12911-023-02179-3.

[16] M. N. Hasan, M. A. Hossain, and M. A. Rahman, "An ensemble based lightweight deep learning model for the prediction of cardiovascular diseases from electrocardiogram images," *Engineering Applications of Artificial Intelligence*, vol. 141, Feb. 2025, Art. no. 109782, https://doi.org/10.1016/j.engappai.2024.109782.