

A Novel Hybrid Feature Selection Framework with Dimensionality Reduction for Early Cardiovascular Disease Detection

G. Muthuselvi* and J. Jebamalar
Tamilselvi

Department of Computer Science,
SRM Institute of Science and
Technology, Ramapuram, Chennai
600089, India

*Email: mselvi.data@gmail.com

Received for publication
May 14, 2026.

Abstract

Cardiovascular disease (CVD) is one of the leading causes of mortality worldwide, necessitating the development of accurate and efficient predictive models for early diagnosis and intervention. However, the high dimensionality, redundancy, and heterogeneity of clinical data often limit the performance of conventional machine learning (ML) approaches. To address these challenges, this study proposes a hybrid feature selection with dimensionality reduction framework for CVD detection (HFSDR-CVD). The proposed framework integrates principal component analysis for dimensionality reduction, mutual information-based feature ranking for relevance assessment, and recursive feature elimination for optimal feature subset selection. The optimized feature space is subsequently used to train and evaluate multiple ML classifiers, including logistic regression, support vector machine, random forest, gradient boosting, and extreme gradient boosting (XGBoost). Experimental evaluation demonstrates that the proposed framework effectively reduces the original feature set from 13 to 6, achieving a feature reduction rate of 53.85% while preserving critical diagnostic information. Among the evaluated classifiers, XGBoost achieved the highest performance with an accuracy of 98.31%, precision of 98.06%, recall of 97.95%, F1-score of 98.00%, and ROC-AUC of 99.02%. Comparative analysis with recent state-of-the-art CVD prediction methods confirms the superiority of the proposed HFSDR-CVD framework in terms of predictive accuracy, feature optimization, and computational efficiency.

Keywords

cardiovascular disease prediction, machine learning, hybrid feature selection, dimensionality reduction, principal component analysis, random forest, early disease detection, healthcare analytics

1. Introduction

Early diagnosis of cardiovascular disease (CVD) enables healthcare professionals to identify high-risk individuals before severe complications occur. Traditional diagnostic procedures typically involve clinical examinations, laboratory investigations, electrocardiograms, echocardiography, angiography, and

physician-based assessments. Although these methods provide valuable clinical insights, they are often time-consuming, expensive and dependent on expert interpretation. Furthermore, CVDs frequently develop gradually over time, making it challenging to detect subtle patterns and risk factors through conventional diagnostic approaches alone [1–3]. The growing volume of electronic health records, laboratory

measurements, demographic information, and physiological parameters presents an opportunity to utilize advanced computational techniques to identify hidden relationships within medical data. Therefore, intelligent predictive systems capable of analyzing complex healthcare datasets have become increasingly important in supporting clinicians and facilitating timely interventions [4,5].

The rapid advancement of artificial intelligence (AI) and machine learning (ML) technologies has transformed numerous sectors, including health care, by enabling data-driven decision-making and predictive analytics. ML algorithms possess the capability to learn intricate patterns from historical medical records and generate accurate predictions regarding disease occurrence and progression. In CVD prediction, ML techniques such as logistic regression (LR), Decision Trees, random forests (RFs), support vector machines (SVMs), gradient boosting (GB) Machines, and extreme gradient boosting (XGBoost) have demonstrated remarkable potential in identifying patients at risk of developing cardiovascular complications [6–8]. These algorithms can process large volumes of heterogeneous clinical data and uncover complex non-linear relationships among risk factors that may not be readily apparent through traditional statistical methods. As a result, ML-based diagnostic systems have become valuable tools for enhancing clinical decision support and improving healthcare outcomes.

Despite significant progress in ML-based CVD prediction, several challenges continue to hinder the development of highly accurate and efficient predictive models. One of the primary challenges is the high dimensionality of healthcare datasets. Modern medical databases often contain numerous clinical attributes, laboratory measurements, physiological indicators, lifestyle factors, and demographic variables. While these features collectively provide comprehensive patient information, not all attributes contribute equally to disease prediction [9–12]. Many variables may be redundant, irrelevant, noisy, or highly correlated with other features. The inclusion of such unnecessary information can negatively impact model performance by increasing computational complexity, introducing overfitting, and reducing generalization capability. Consequently, predictive models trained on high-dimensional datasets may exhibit suboptimal performance despite the availability of extensive data [13–15].

Another critical issue associated with high-dimensional medical data is the curse of dimensionality. As the number of features increases, the volume of the feature space expands exponentially, making it

increasingly difficult for ML algorithms to identify meaningful patterns and relationships. This phenomenon often leads to increased training times, higher memory requirements, and degraded predictive accuracy. Moreover, redundant and irrelevant attributes may obscure the contribution of clinically significant risk factors, thereby reducing model interpretability. Since healthcare applications require reliable and explainable predictions, addressing the challenges associated with high-dimensional data has become a key research priority in medical data analytics.

Feature optimization techniques have emerged as effective solutions for overcoming the limitations of high-dimensional healthcare datasets. Feature selection aims to identify the most informative attributes while eliminating irrelevant and redundant variables. By reducing the number of input features, feature selection enhances model efficiency, decreases computational cost, and improves classification performance. Existing feature selection approaches are generally categorized into filter methods, wrapper methods, and embedded methods. Filter methods evaluate feature relevance independently of ML models, whereas wrapper methods utilize classifier performance to assess feature subsets. Embedded methods perform feature selection during model training. Although these techniques have demonstrated promising results individually, each possesses inherent limitations. Filter methods may overlook feature interactions, while wrapper methods often require substantial computational resources. Consequently, combining multiple feature selection strategies within a hybrid framework can leverage the strengths of different approaches and achieve superior optimization performance.

In addition to feature selection, dimensionality reduction techniques play a vital role in managing high-dimensional medical data. Dimensionality reduction transforms the original feature space into a compact representation while preserving the most significant information contained within the dataset. Principal component analysis (PCA) is one of the most widely used dimensionality reduction techniques due to its ability to capture maximum variance and eliminate multicollinearity among features. By reducing data complexity and removing redundant information, dimensionality reduction contributes to improved model stability, faster training, and enhanced predictive performance. However, dimensionality reduction alone may not always identify the most clinically relevant attributes. Therefore, integrating dimensionality reduction with feature selection provides a comprehensive strategy for constructing highly informative feature subsets [16–20].

The motivation for this research stems from the need to develop an intelligent and efficient CVD prediction framework that is capable of addressing the limitations of existing ML approaches. While numerous studies have employed ML algorithms for CVD diagnosis, many rely on either standalone feature selection or dimensionality reduction methods. Such approaches may fail to fully exploit the complementary advantages offered by combined feature optimization strategies. Furthermore, existing models often prioritize prediction accuracy without adequately considering computational efficiency and feature relevance. To bridge these gaps, this study proposes a Novel Hybrid Feature Selection Framework with Dimensionality Reduction for Early CVD Detection. The proposed framework integrates PCA for dimensionality reduction with a hybrid feature selection mechanism that combines filter-based and wrapper-based optimization strategies. The resulting optimized feature subset is subsequently evaluated using multiple ML classifiers to identify the most effective predictive model.

By integrating dimensionality reduction and hybrid feature optimization within a unified framework, the proposed approach seeks to enhance predictive accuracy, reduce computational overhead, improve model interpretability, and facilitate reliable clinical decision support. The framework is designed to assist healthcare practitioners in identifying high-risk patients at an early stage, thereby enabling timely intervention and reducing the burden associated with CVDs. Through comprehensive experimentation and comparative analysis, this research aims to demonstrate the effectiveness of feature optimization techniques in improving ML-based CVD prediction systems.

a. Contributions of the proposed work

The major contributions of this study are summarized as follows:

1. A hybrid feature selection methodology combining filter-based and wrapper-based techniques is developed to identify the most informative cardiovascular risk factors while eliminating redundant and irrelevant features.
2. PCA is incorporated with hybrid feature selection to effectively address the challenges associated with high-dimensional medical datasets and improve data representation.
3. The optimized feature subset is evaluated using various ML classifiers, including LR, SVM, RF, GB, and XGBoost, to determine the most effective prediction model.
4. The proposed framework enhances CVD prediction accuracy while simultaneously reducing feature dimensionality, training time, memory consumption, and overall computational cost.
5. The framework provides a reliable and efficient decision-support mechanism capable of assisting healthcare professionals in identifying CVD risks at an early stage, thereby enabling timely preventive interventions and improved patient outcomes.

II. Literature Review

CVD remains one of the most prevalent and life-threatening medical conditions worldwide, prompting extensive research into intelligent prediction systems capable of facilitating early diagnosis and intervention. Recent advancements in AI, ML, deep learning (DL), feature selection, and ensemble learning have significantly enhanced the accuracy and reliability of CVD prediction models. Despite these developments, challenges associated with high-dimensional medical data, feature redundancy, computational complexity, and model interpretability continue to motivate further investigation. This section reviews recent studies related to CVD prediction, feature optimization, dimensionality reduction, and hybrid ML frameworks.

Recent studies have demonstrated the effectiveness of ML algorithms in identifying CVD risk factors from clinical datasets. Raman et al. [14] proposed an enhanced CVD prediction framework that integrated advanced ML algorithms with hybrid feature selection techniques. Their study demonstrated that the selection of relevant clinical attributes significantly improved classification performance by reducing the influence of redundant and noisy features. The authors highlighted the importance of combining multiple feature selection approaches to achieve better predictive accuracy compared with conventional methods.

Ahmed and Husien [3] presented a comprehensive review of hybrid ML approaches for heart disease prediction. Their study examined various combinations of ML classifiers and optimization techniques, emphasizing that hybrid frameworks generally outperform standalone classifiers due to their ability to capture complex relationships among clinical variables. The review identified feature optimization as one of the most critical factors influencing prediction accuracy and generalization performance.

Similarly, Abdullahi et al. [2] investigated ML-based CVD prediction using advanced feature selection methodologies. Their findings revealed that eliminating

irrelevant attributes improved classifier efficiency and reduced overfitting. The study further demonstrated that optimized feature subsets enhanced the predictive capabilities of traditional ML models such as LR, SVMs, and RFs.

The increasing adoption of ensemble learning techniques has further improved CVD diagnosis. Al-Sayed et al. [4] proposed a collaborative clustering and ensemble learning framework for heart disease prediction. Their methodology combined clustering-based preprocessing with ensemble classifiers to improve disease classification performance. Experimental results indicated that ensemble models effectively handled data variability and produced more reliable predictions than individual classifiers.

Likewise, Mandula and Vijaya Kumar [9] introduced a two-tier heart disease detection framework integrating feature selection with ensemble learning. Their approach employed an adaptive learning artificial network and ensemble tree-based adaptive boosting decision forest models to improve classification accuracy. The study demonstrated that combining feature optimization with ensemble learning significantly enhanced predictive performance while maintaining computational efficiency.

Further advancements in ensemble-based CVD prediction were reported by Khan et al. [7], who developed EnsCVDD-Net and BICVDD-Net, two novel ensemble and blending-based CVD detection networks. Their results demonstrated superior accuracy, precision, and recall compared with traditional ML approaches. The study highlighted the capability of ensemble architectures to capture non-linear relationships among clinical features and improve disease classification reliability.

DL has also emerged as a powerful tool for CVD diagnosis due to its ability to automatically learn hierarchical feature representations from complex medical datasets. Rao et al. (2024) proposed AttGRU-HMSI, a hybrid DL framework integrating attention mechanisms with gated recurrent units for heart disease diagnosis. The model effectively captured temporal and non-linear dependencies within clinical data, resulting in substantial improvements in diagnostic accuracy. The incorporation of attention mechanisms further enhanced model interpretability by identifying influential risk factors.

Jayasudha et al. [5] developed a hybrid optimization-enabled DL ensemble classification framework for heart disease detection. Their approach integrated optimization algorithms with DL classifiers to enhance feature representation and classification performance. Experimental findings demonstrated

superior accuracy and robustness compared with conventional DL models, highlighting the benefits of optimization-driven learning architectures.

Similarly, Kallimani et al. [6] proposed a hybrid DL-based heart disease detection framework incorporating feature selection mechanisms within an e-healthcare environment. The study emphasized the significance of selecting informative clinical features prior to model training. Results indicated that hybrid feature selection considerably improved classification performance while reducing computational overhead.

Subramani et al. [17] explored the integration of ML and DL techniques for CVD prediction. Their study revealed that combining traditional ML models with deep neural networks could improve disease detection capabilities by leveraging complementary strengths of both approaches. The authors concluded that hybrid architectures offer promising solutions for real-world healthcare applications requiring high predictive accuracy and scalability.

Recent research has also focused on representation learning and advanced ensemble frameworks for CVD diagnosis. Abdellatif et al. [1] introduced a computational heart disease detection system based on conditional variational autoencoders and stacked ensemble learning. The proposed framework generated latent feature representations that enhanced classifier performance. Experimental results confirmed the effectiveness of combining deep feature extraction with ensemble classification for accurate disease prediction.

Mienye and Sun [10] proposed a particle swarm optimization (PSO)-based stacked sparse autoencoder for heart disease prediction. Their approach utilized optimization techniques to improve feature learning and parameter tuning. The study reported substantial gains in prediction accuracy and demonstrated the potential of evolutionary optimization algorithms in healthcare analytics.

Razzaque and Badholia [15] presented a framework combining PCA for feature extraction with Modified PSO (MPSO) for feature selection in gene expression-based medical data classification. Their results demonstrated that dimensionality reduction followed by optimization-driven feature selection significantly enhanced classification accuracy while reducing computational complexity. The study highlights the complementary benefits of integrating dimensionality reduction and feature optimization techniques.

The broader landscape of AI-driven cardiovascular health care has also witnessed significant developments. Paul and Masood [12] conducted

a comprehensive survey of predictive methods for CVD, reviewing ML, DL, and ensemble learning approaches. Their analysis revealed that hybrid models generally achieve higher diagnostic accuracy than standalone techniques. The authors emphasized the growing importance of explainable AI, feature optimization, and multimodal learning in future CVD prediction systems.

Milosevic et al. [11] explored applications of AI in multimodal cardiovascular imaging. Their work demonstrated that AI-based models can effectively analyze imaging data from multiple modalities to support CVD diagnosis. Although imaging-based approaches differ from structured clinical data analytics, the study underscored AI's transformative role in improving diagnostic precision and clinical decision-making.

Apart from healthcare-specific studies, advancements in ML architectures have contributed significantly to predictive analytics. Santosh Kumar and Bharathi [16] proposed ConvTransFormer, a hybrid framework integrating convolutional neural networks and transformers for signal processing applications. Their work highlighted the advantages of combining multiple learning paradigms to improve feature extraction and classification performance. Similar hybrid principles have increasingly been adopted in healthcare prediction models.

III. Proposed HFSDR-CVD Framework

a. Overview of the proposed framework

The proposed Hybrid Feature Selection with Dimensionality Reduction Framework for CVD Detection (HFSDR-CVD) is designed to address two major challenges commonly observed in CVD prediction systems: the presence of high-dimensional clinical data and the existence of redundant or irrelevant attributes that influence classifier performance negatively. Although contemporary ML algorithms have demonstrated promising results in CVD diagnosis, their predictive capability is often limited by noisy feature spaces, multicollinearity among clinical variables, and increased computational complexity. To overcome these limitations, the proposed framework integrates dimensionality reduction and hybrid feature optimization into a unified predictive architecture.

The overall framework consists of four interconnected modules: data preprocessing, principal component-based dimensionality reduction, hybrid feature optimization, and ML-based disease prediction.

Initially, raw cardiovascular data undergo preprocessing operations to eliminate inconsistencies and standardize feature distributions. Subsequently, PCA is employed to reduce feature dimensionality by transforming correlated variables into a reduced set of orthogonal components. While PCA effectively removes multicollinearity and preserves the majority of information contained in the dataset, certain transformed components may still contribute minimally toward disease classification. Therefore, a hybrid feature selection mechanism is introduced to further identify the most discriminative features.

The proposed hybrid feature optimization strategy combines filter-based feature relevance analysis and wrapper-based subset optimization. The filter stage evaluates feature importance using mutual information (MI) and correlation analysis, allowing the framework to eliminate weakly associated variables at an early stage. The remaining candidate features are subsequently processed through recursive feature elimination (RFE), which iteratively removes less informative features based on classifier performance. This two-stage optimization process ensures that only highly informative and non-redundant attributes are retained for model training.

Finally, the optimized feature subset is supplied to multiple ML classifiers, including LR, SVM, RF, GB, and XGBoost. Comparative evaluation across multiple classifiers enables identification of the most effective predictive model for CVD detection. The complete architecture of the proposed HFSDR-CVD framework is illustrated in Figure 1.

b. Data preprocessing and feature normalization

The effectiveness of ML-based CVD prediction largely depends on the quality of the input data. Clinical datasets often contain missing observations, noisy measurements, outliers, and features with heterogeneous numerical scales. These issues can adversely affect the learning capability of classifiers, resulting in reduced prediction accuracy and poor generalization performance. Therefore, a comprehensive preprocessing stage is incorporated into the proposed HFSDR-CVD framework to improve data quality and generate a consistent feature representation suitable for subsequent dimensionality reduction and feature optimization.

Let the CVD dataset be represented as

$$D = \{(x_i, y_i)\}_{i=1}^N$$

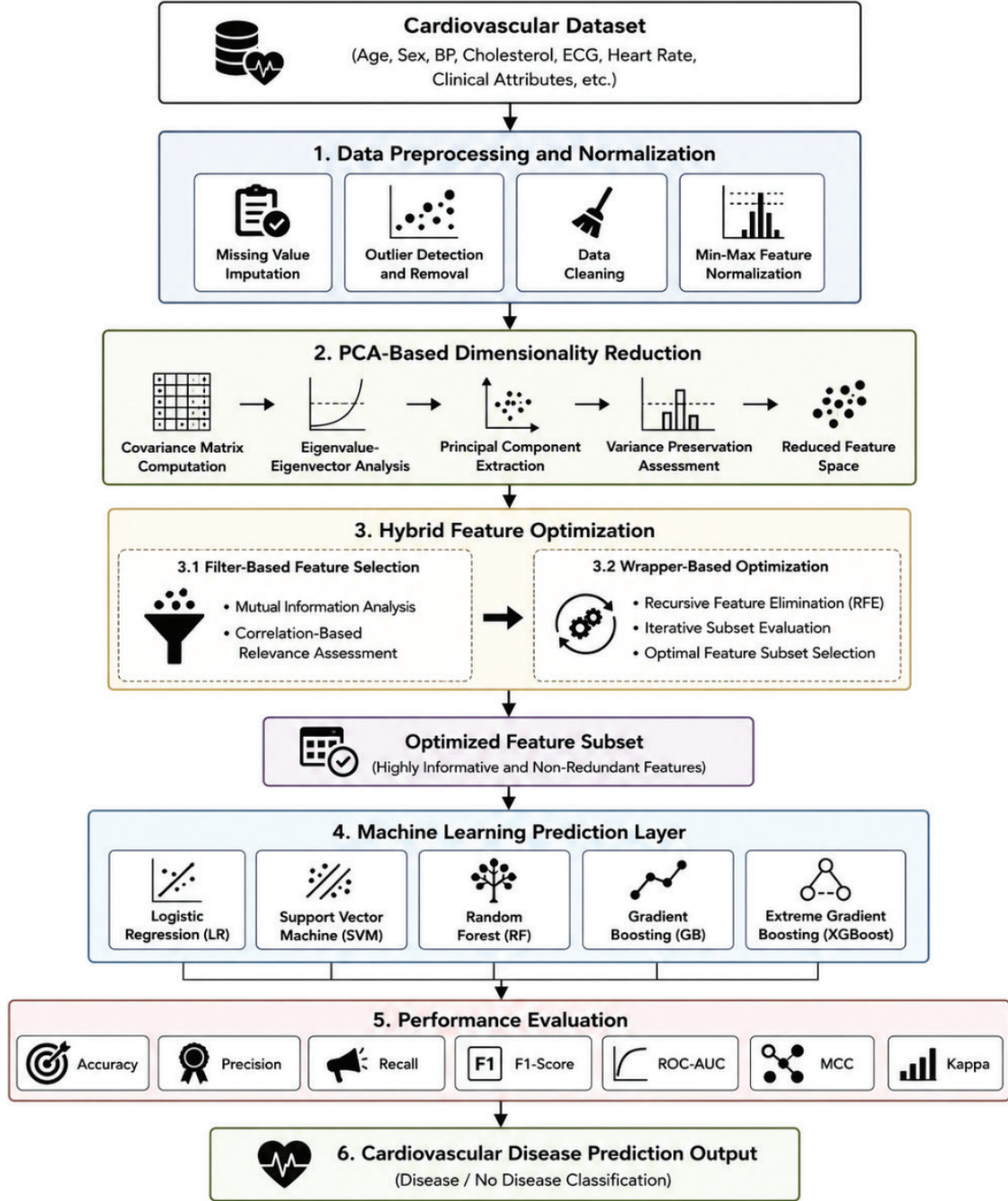


Figure 1: Model architecture. GB, gradient boosting; LR, logistic regression; PCA, principal component analysis; RF, random forest; RFE, recursive feature elimination; SVM, support vector machine; XGBoost, extreme gradient boosting.

where N denotes the total number of patient records, $x_i = [x_{i1}, x_{i2}, \dots, x_{im}]$ represents the feature vector containing m clinical attributes, and y_i denotes the corresponding class label indicating the presence or absence of CVD.

b.i Missing value treatment

Medical datasets frequently contain incomplete records due to unavailable laboratory measurements, data entry errors, or inconsistencies during

data collection. The presence of missing values may introduce bias and negatively influence model performance. To ensure data completeness, missing numerical values are replaced through mean-value imputation.

For a feature X_j , the mean value is calculated as

$$\widehat{X}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij}$$

where n_j denotes the number of valid observations available for the j^{th} feature.

The missing entries are subsequently replaced as

$$X_{ij} = \begin{cases} X_{ij} & \text{if observed} \\ \widehat{X}_j & \text{if missing} \end{cases}$$

This process preserves the statistical characteristics of the dataset while eliminating incomplete records.

b.ii Outlier detection and removal

Outliers are abnormal observations that significantly deviate from the majority of data points. Such extreme values can distort feature distributions and adversely affect classifier learning. Therefore, outlier detection is performed using the interquartile range (IQR) method.

The IQR is computed as

$$IQR = Q_3 - Q_1$$

where Q_1 and Q_3 denote the first and third quartiles, respectively.

The lower and upper boundaries for identifying outliers are determined as

$$LB = Q_1 - 1.5(IQR)$$

$$UB = Q_3 + 1.5(IQR)$$

A data point is considered an outlier if

$$X < LB \text{ or } X > UB$$

Detected outliers are removed from the dataset to improve data reliability and reduce noise during model training.

b.iii Feature standardization

Clinical attributes such as cholesterol level, blood pressure, age, fasting blood sugar, and heart rate are measured on different numerical scales. Features with larger magnitudes may dominate the learning process and negatively influence model convergence. To address this issue, feature standardization is applied.

The standardized value of a feature is computed using z-score normalization:

$$z_i = \frac{X_i - \mu}{\sigma}$$

where μ represents the mean value of the feature and σ denotes its standard deviation.

The mean is calculated as

$$\mu = \frac{1}{N} \sum_{i=1}^N X_i$$

while the standard deviation is given by

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (X_i - \mu)^2}$$

Standardization transforms all features to a common scale with zero mean and unit variance, thereby improving numerical stability and classifier performance.

b.iv Min-max normalization

Following standardization, Min-Max normalization is employed to map all feature values into the interval [0,1]. This transformation enhances the efficiency of optimization-based learning algorithms and prevents bias toward features with larger ranges.

The normalized feature value is computed as

$$x' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

where x denotes the original feature value, and $x_{\min}$ and $x_{\max}$ represent the minimum and maximum values of the feature, respectively.

The transformed feature satisfies

$$0 \leq x' \leq 1$$

for all observations.

b.v Construction of the preprocessed dataset

After missing value treatment, outlier removal, standardization, and normalization, the processed dataset is represented as

$$D_p = \begin{bmatrix} x'_{11} & 0x'_{12} & \dots & x'_{1m} \\ x'_{21} & x'_{22} & & x'_{2m} \\ x'_{N1} & x'_{N2} & & x'_{Nm} \end{bmatrix}$$

where D_p denotes the final normalized feature matrix containing N patient records and m preprocessed attributes.

The preprocessing stage produces a clean, standardized, and normalized dataset that minimizes noise, eliminates inconsistencies, and improves feature quality. This optimized dataset serves as the input for the subsequent PCA-based dimensionality reduction and hybrid feature selection modules within the proposed HFSDR-CVD framework.

c. Principal component-based dimensionality reduction

The CVD dataset contains multiple clinical and physiological attributes that often exhibit strong correlations and redundant information. The presence of such high-dimensional features increases computational complexity, prolongs training time, and may lead to overfitting in ML models. Furthermore, redundant variables can obscure the contribution of significant risk factors and negatively affect the predictive capability of classifiers. Therefore, dimensionality reduction is incorporated into the proposed HFSDR-CVD framework to generate a compact and informative feature representation while preserving the essential characteristics of the original dataset.

Among various dimensionality reduction techniques, PCA is selected due to its ability to transform correlated variables into a reduced set of uncorrelated components while retaining the maximum variance contained in the data. PCA reduces feature dimensionality by projecting the original dataset onto a new orthogonal feature space, thereby eliminating redundancy and improving computational efficiency.

c.i PCA

Let the preprocessed dataset be represented by

$$D_p = \{x_1, x_2, \dots, x_N\}$$

where N denotes the number of patient records and each observation contains m normalized clinical attributes.

The first step in PCA is constructing the covariance matrix of the normalized dataset. The covariance matrix captures the linear relationships among features and is computed as

$$C = \frac{1}{N-1} X^T X$$

where X represents the normalized feature matrix, and C denotes the covariance matrix.

Subsequently, eigenvalue decomposition is performed on the covariance matrix to obtain the eigenvectors and eigenvalues:

$$Cv_i = \lambda_i v_i$$

where v_i represents the eigenvector corresponding to the i^{th} principal component and λ_i denotes its associated eigenvalue.

The eigenvectors define the directions of maximum variance, whereas the eigenvalues quantify the amount of variance captured by each principal component. The principal components are arranged in descending order according to their eigenvalues such that

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$$

The transformed feature space is obtained through

$$Z = XV$$

where V represents the matrix of selected eigenvectors and Z denotes the projected low-dimensional feature space.

c.ii Variance preservation analysis

An important objective of PCA is to retain the maximum amount of information while reducing dimensionality. Since each eigenvalue represents the variance captured by its corresponding principal component, the proportion of variance explained by the i^{th} component is calculated as

$$EV_i = \frac{\lambda_i}{\sum_{j=1}^m \lambda_j}$$

where EV_i denotes the explained variance ratio.

To evaluate the total information preserved by multiple components, the cumulative explained variance is computed as

$$CEV_k = \sum_{i=1}^k EV_i$$

where k represents the number of retained principal components.

The cumulative variance analysis provides a quantitative measure of information retention and assists in determining the optimal dimensionality of the transformed feature space. A higher cumulative variance indicates better preservation of the original dataset characteristics.

c.iii Optimal component selection

After calculating the cumulative explained variance, an optimal subset of principal components is selected to balance information preservation and dimensionality reduction. Components are retained until the cumulative variance exceeds a predefined threshold τ , typically ranging from 90% to 95%.

The component selection criterion is expressed as

$$k = \min \left\{ k : \sum_{i=1}^k \lambda_i \geq \tau \sum_{j=1}^m \lambda_j \right\}$$

where k denotes the minimum number of principal components required to preserve the desired variance threshold.

The resulting reduced feature matrix is represented as

$$Z_k = XV_k$$

where V_k contains the selected k principal eigen vectors.

By retaining only the most informative components, PCA effectively removes feature redundancy, reduces multicollinearity, and lowers computational complexity while preserving the dominant information contained within the cardiovascular dataset. The reduced feature matrix generated through PCA serves as the input for the subsequent hybrid feature selection module, enabling further refinement of the feature space and improving CVD prediction performance.

d. Hybrid feature selection and feature optimization

Although PCA effectively reduces feature dimensionality and eliminates multicollinearity, the transformed feature space may still contain components that contribute minimally toward CVD classification. Furthermore, PCA primarily focuses on variance

preservation and does not explicitly evaluate the predictive relevance of features with respect to the target class. Consequently, an additional feature optimization stage is required to identify the most informative attributes that significantly influence CVD prediction.

To address this challenge, the proposed HFSDR-CVD framework incorporates a hybrid feature selection (HFS) strategy that combines the advantages of both filter-based and wrapper-based feature selection methods. The integration of these complementary approaches enables the framework to eliminate irrelevant features at an early stage while simultaneously identifying the feature subset that maximizes classifier performance. The hybrid methodology consists of three sequential steps: (i) filter-based feature ranking, (ii) wrapper-based feature optimization, and (iii) generation of the optimal feature subset.

d.i Filter-based feature ranking

The first stage of the proposed hybrid feature selection framework employs a filter-based mechanism to evaluate the relevance of each feature independently of the classification model. Filter methods are computationally efficient and capable of eliminating weakly informative attributes quickly before the optimization stage.

In this study, feature relevance is assessed using MI and Correlation Analysis.

d.ii Mutual information-based relevance analysis

MI measures the degree of dependency between a feature and the target class. Features with higher MI values provide greater predictive information regarding CVD occurrence.

The MI between feature X and class label Y is computed as

$$MI(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \left(\frac{p(x, y)}{p(x) p(y)} \right)$$

where:

- $p(x, y)$ denotes the joint probability distribution,
- $p(x)$ represents the marginal probability of feature X , and
- $p(y)$ represents the marginal probability of class label Y .

A larger MI score indicates a stronger association between the feature and CVD prediction.

d.iii Correlation-based feature analysis

To further identify redundant attributes, Pearson correlation coefficients are computed between pairs of features.

The correlation coefficient between two features X_i and X_j is given by

$$\rho_{ij} = \frac{\sum_{k=1}^N (X_{ik} - \hat{X}_i)(X_{jk} - \hat{X}_j)}{\sqrt{\sum_{k=1}^N (X_{ik} - \hat{X}_i)^2 \sum_{k=1}^N (X_{jk} - \hat{X}_j)^2}}$$

where:

- $\hat{X}_i$ and $\hat{X}_j$ denote feature means and
- N represents the total number of samples.

Features exhibiting strong correlation ($|\rho| > \delta$) are considered redundant, and only the most informative attribute is retained.

The combined filter relevance score is defined as

$$FR_i = \alpha \cdot MI_i + (1 - \alpha)(1 - |\rho_i|)$$

where:

- FR_i denotes the filter ranking score,
- MI_i represents MI,
- ρ_i represents correlation, and
- α represents the weighting factor.

Features are ranked according to R_i , and only the highest-ranked attributes are forwarded to the wrapper stage.

d.iv Wrapper-based optimization using RFE

Although filter methods efficiently remove irrelevant features, they do not consider classifier-specific feature interactions. Therefore, a wrapper-based optimization stage is employed to identify the subset that maximizes prediction performance.

The proposed framework utilizes RFE as the wrapper optimization technique. RFE iteratively trains a ML classifier and evaluates the contribution of each feature toward classification accuracy.

Let the ranked feature set be represented as

$$F = \{f_1, f_2, \dots, f_n\}$$

where n denotes the number of retained features after filter ranking.

At each iteration, feature importance scores are calculated:

$$I(f_i) = w_i$$

where w_i denotes the importance weight assigned to feature f_i

The least significant feature is removed according to

$$f_{\text{remove}} = \arg \min I(f_i)$$

The feature subset is updated as

$$F_{t+1} = F_t - f_{\text{remove}}$$

where:

- F_t denotes the current feature subset and
- F_{t+1} denotes the reduced subset after elimination.

The elimination process continues until the optimal subset yielding maximum classifier performance is obtained.

The objective function of RFE can be expressed as

$$F^* = \arg \max Acc(F_k)$$

where:

F^* denotes the optimal feature subset and

$Acc(F_k)$ represents classification accuracy obtained using subset F_k .

This procedure ensures that only the most discriminative features contributing to CVD prediction are retained.

d.v Generation of the optimal feature subset

The final stage combines the outputs of filter ranking and wrapper optimization to generate a compact yet highly informative feature subset.

The optimal feature subset can be represented as

$$F_{\text{opt}} = \{f_1, f_2, \dots, f_k\}$$

where:

$$k \ll n$$

and k denotes the number of selected features after optimization.

The dimensionality reduction ratio achieved by the proposed framework is calculated as

$$DR = \left(1 - \frac{k}{m}\right) \times 100$$

where:

- m is the total number of original features and
- k is the number of optimized features.

To evaluate the effectiveness of the selected feature subset, a feature quality score is defined as

$$FQ = \frac{\sum_{i=1}^k MI_i}{k}$$

where FQ represents the average information contribution of the selected features.

The resulting optimized feature matrix is represented as

$$X_{opt} = [X_{f1}, X_{f2}, \dots, X_{fk}]$$

which contains only the most informative attributes relevant to CVD diagnosis.

The optimized feature matrix X_{opt} generated by the dimensionality reduction and hybrid feature selection modules is subsequently used for CVD prediction. Rather than relying on a single classifier, the proposed HFSDR-CVD framework evaluates multiple ML models with complementary learning characteristics. This strategy enables a rigorous comparison between linear, kernel-based, bagging-based, and boosting-based classifiers and facilitates identification of the model that best exploits the optimized cardiovascular feature space.

Let the final training dataset be represented as

$$D_{opt} = \{(x_i, y_i)\}_{i=1}^N, x_i \in R^k, y_i \in \{0, 1\}$$

where k denotes the number of selected features after PCA and hybrid feature optimization. The prediction objective is to learn a mapping

$$f: R^k \rightarrow \{0, 1\}$$

that minimizes classification error on unseen patient records.

e. ML-based CVD prediction

e.i LR

LR is employed as a baseline probabilistic classifier because of its interpretability and strong performance on structured clinical data. The model estimates the probability that a patient belongs to the CVD class using the sigmoid function

$$P(Y = 1 | x) = \sigma(z) = \frac{1}{1 + e^{-z}}$$

$$z = \beta_0 + \sum_{j=1}^k \beta_j x_j$$

where β_0 is the intercept and β_j are feature coefficients. Model parameters are estimated by minimizing the negative log-likelihood (cross-entropy) loss

$$L_{LR} = -\sum_{i=1}^N [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

with $p_i = P(Y = 1 | x_i)$

e.ii SVM

SVM constructs a separating hyperplane that maximizes the margin between disease and non-disease classes. For linearly separable data, the primal optimization problem is

$$\min_{w, b} \frac{1}{2} \|w\|^2$$

$$y_i(w^T x_i + b) \geq 1, i = 1, \dots, N$$

For non-separable data, slack variables and a penalty parameter C are introduced

$$\min_{w, b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi_i$$

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0$$

Kernel functions can be employed when non-linear decision boundaries are required.

e.iii RF

RF is a bagging-based ensemble method that combines multiple decision trees trained on bootstrap

samples. Each tree performs classification independently, and the final prediction is obtained through majority voting

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \dots, T_M(x)\}$$

where T_m denotes the prediction of the m^{th} tree and MMM is the total number of trees. RF improves robustness, reduces variance, and is relatively resistant to overfitting.

e.iv GB

GB constructs an additive ensemble by sequentially fitting weak learners to the residual errors of previous learners. The ensemble prediction at iteration m is

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

where $h_m(x)$ is the weak learner trained on the current residuals and γ_m is the learning rate or step size. The model minimizes a differentiable loss function by performing gradient descent in function space.

e.v XGBoost

XGBoost extends GB with regularization, second-order optimization, shrinkage, and efficient tree construction. The objective function at iteration t is

$$L^{(t)} = \sum_{i=1}^N l(y_i, \hat{y}_i(t)) + \sum_{k=1}^K \Omega(f_k)$$

where $l(\cdot)$ denotes the classification loss, f_k represents the k^{th} regression tree, and $\Omega(f_k)$ is a regularization term controlling model complexity. XGBoost is included because of its strong predictive performance on tabular clinical datasets.

e.vi Training and model selection

All classifiers are trained using the same optimized feature matrix X_{opt} to ensure a fair comparison. Hyperparameters are tuned using cross-validation on the training set, and the best configuration is selected according to validation performance. For a classifier with parameter vector θ , the learning objective can be summarized as

$$\theta^* = \min_{\theta} L(D_{\text{train}}; \theta)$$

where L denotes the model-specific training loss. After optimization, the trained model produces predictions

$$\hat{y}_i = f_{\theta^*}(x_i)$$

for previously unseen patient records.

e.vii Computational complexity considerations

One of the motivations for the proposed HFSDR-CVD framework is to reduce computational burden through feature optimization. If the original dataset contains mmm features and the optimized subset contains k features ($k \ll m$), the dimensionality reduction ratio is

$$DR = \left(1 - \frac{k}{m}\right) \times 100$$

Reducing the effective feature dimensionality lowers memory usage, decreases training time, and improves scalability for large clinical datasets while preserving predictive information relevant to CVD detection.

f. Proposed HFSDR-CVD algorithm and computational complexity analysis

The proposed HFSDR-CVD integrates PCA, MI-based feature ranking, RFE, and ML classification within a unified optimization framework. The primary objective is to maximize predictive performance while simultaneously reducing feature redundancy, dimensionality, and computational cost. Unlike conventional CVD prediction systems that rely solely on dimensionality reduction or standalone feature selection techniques, the proposed framework employs a sequential optimization strategy that preserves informative clinical characteristics and eliminates irrelevant attributes before classification.

The overall optimization process begins with pre-processing and normalization of cardiovascular clinical data. PCA is subsequently applied to transform the original feature space into a lower-dimensional representation while retaining maximum variance. The transformed features are then evaluated using MI to measure their relevance with respect to disease classification. Features with low relevance scores are discarded, and the remaining candidate features are

subjected to RFE to identify the optimal subset that maximizes classification performance. Finally, the optimized feature set is used for training multiple ML classifiers, and the best-performing model is selected for CVD prediction.

f.i Optimization objective function

The proposed hybrid feature selection framework seeks to identify an optimal feature subset that maximizes feature relevance and classification accuracy while minimizing feature redundancy. The optimization objective is formulated as

$$J(F) = \alpha R(F) + \beta A(F) - \gamma D(F)$$

where

- $J(F)$ denotes the overall feature optimization objective,
- $R(F)$ represents feature relevance,
- $A(F)$ denotes classification accuracy,
- $D(F)$ represents feature redundancy, and
- α , β , and γ are weighting coefficients satisfying

$$\alpha + \beta + \gamma = 1$$

The relevance component is calculated using the average MI score of the selected features:

$$R(F) = \frac{1}{k} \sum_{i=1}^k MI(f_i, Y)$$

where k denotes the number of selected features and $MI(f_i, Y)$ represents the MI between feature f_i and the target class Y .

The redundancy term is computed using pairwise feature correlations:

$$D(F) = \frac{2}{k(k-1)} \sum_{i=1}^k \sum_{j=i+1}^k |\rho_{ij}|$$

where ρ_{ij} denotes the correlation coefficient between selected features f_i and f_j .

The classification accuracy term is defined as

$$A(F) = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

The optimal feature subset is obtained by maximizing the objective function:

$$F^* = \arg \max_F J(F)$$

This formulation ensures that the selected features are highly informative, minimally redundant, and capable of improving CVD prediction performance.

Algorithm 1: HFSDR-CVD

Algorithm 1: HFSDR-CVD Framework for Cardiovascular Disease Prediction

Input: Cardiovascular Disease Dataset D

Output: Optimized Feature Subset F_{opt} , Best Prediction Model M^* , Predicted Class Labels Y_{pred}

Begin

1. Load cardiovascular disease dataset D .
2. Perform data preprocessing:
 - Handle missing values.
 - Remove outliers.
 - Normalize feature values.
3. Generate preprocessed dataset D_p .
4. Apply Principal Component Analysis (PCA).
5. Compute principal components and variance contribution.
6. Retain significant components satisfying the variance threshold.
7. Generate reduced feature space Z .
8. Compute Mutual Information scores for all features.
9. Rank features according to relevance.
10. Apply correlation analysis and eliminate redundant features.
11. Generate candidate feature subset F_c .
12. Apply Recursive Feature Elimination (RFE).
13. Identify optimized feature subset F_{opt} .
14. Train Logistic Regression classifier.
15. Train Support Vector Machine classifier.
16. Train Random Forest classifier.
17. Train Gradient Boosting classifier.
18. Train XGBoost classifier.
19. Evaluate all classifiers using predefined performance metrics.
20. Select the classifier with the highest overall performance as M^* .
21. Generate cardiovascular disease predictions Y_{pred} .

Return (F_{opt}, M^*, Y_{pred})

End

f.ii Computational complexity analysis

The computational efficiency of the proposed HFSDR-CVD framework is evaluated by analyzing the complexity of its major processing stages. Let N represent the number of patient records, m denote the original number of features, and k represent the optimized feature subset size after dimensionality reduction and feature selection.

The PCA stage involves covariance matrix computation and eigenvalue decomposition. The computational complexity of PCA is approximately $O(Nm^2 + m^3)$, where covariance matrix generation contributes $O(Nm^2)$ and eigenvalue decomposition contributes $O(m^3)$.

The MI-based feature ranking stage evaluates feature relevance independently and requires $O(Nm)$ $O(Nm)O(Nm)$ computational time. Correlation analysis between feature pairs contributes approximately $O(m^2)$.

The RFE process iteratively removes features while repeatedly training the learning model. Assuming k candidate features are evaluated, the complexity of RFE is approximately $O(Nk^2)$.

The ML classification stage depends on the underlying classifier. LR exhibits relatively low complexity, while RF, GB, and XGBoost require additional computation due to ensemble construction. However, because the feature dimension is substantially reduced before classification, the computational burden of all classifiers is significantly lower than that of training on the original dataset.

Consequently, the overall computational complexity of the proposed framework can be approximated as

$$O(Nm^2 + m^3 + Nk^2)$$

where $k \ll m$. This reduction in effective feature dimensionality contributes to improved scalability, reduced training time, and enhanced computational efficiency.

f.iii Advantages of the proposed framework

The proposed HFSDR-CVD framework offers several advantages over conventional CVD prediction approaches.

First, the incorporation of PCA effectively reduces feature dimensionality while preserving most of the information contained within the original dataset.

This transformation minimizes multicollinearity and improves model stability.

Second, the hybrid feature selection mechanism combines filter-based and wrapper-based strategies to exploit their complementary strengths. MI rapidly identifies relevant features, while RFE performs fine-grained optimization by considering classifier-specific feature interactions.

Third, the framework significantly reduces feature redundancy and removes irrelevant clinical attributes, resulting in a compact and highly informative feature subset. This optimized representation improves classifier learning capability and enhances prediction accuracy.

Fourth, dimensionality reduction and feature optimization collectively reduce computational complexity, training time, and memory consumption. The framework is therefore suitable for large-scale healthcare datasets and real-time clinical decision-support applications.

Fifth, the comparative evaluation of multiple ML models enables selection of the most suitable classifier for CVD prediction, thereby improving robustness and generalization performance.

Finally, the proposed framework enhances interpretability by identifying the most influential cardiovascular risk factors, facilitating better clinical understanding and supporting evidence-based medical decision-making.

f.iv Framework output and decision-making process

The final HFSDR-CVD framework consists of three major components: the optimized feature subset, the best-performing classification model, and the CVD prediction results.

The optimized feature subset F_{opt} contains only the most informative and non-redundant clinical attributes identified through the combined PCA and hybrid feature selection process. This subset serves as the foundation for accurate and computationally efficient disease prediction.

The classifier exhibiting the highest performance across Accuracy, Precision, Recall, F1-Score, and ROC-AUC is selected as the final prediction model M^* . This model is subsequently utilized to classify unseen patient records into CVD and non-CVD categories.

For a new patient record x_{new} , the trained model generates a prediction.

$$\hat{y} = M^*(x_{new})$$

where $\hat{y} = 1$ indicates the presence of CVD risk and $\hat{y} = 0$ indicates the absence of significant disease risk.

In addition to class labels, the framework can provide probability-based risk estimates that assist healthcare professionals in evaluating disease severity and making informed clinical decisions. By combining feature optimization with ML-based classification, the HFSDR-CVD framework supports reliable early diagnosis and facilitates proactive CVD management.

IV. Results and Analysis

a. Experimental environment

The proposed HFSDR-CVD framework was implemented using Python 3.11 within the Jupyter Notebook environment. Data preprocessing and feature engineering operations were performed using NumPy and Pandas libraries, while PCA, MI

ranking, RFE, LR, SVM, RF, and GB classifiers were implemented using Scikit-learn. XGBoost was implemented using the XGBoost library. The experiments were conducted on an Intel Core i7 processor with 16 GB RAM running the Windows operating system.

b. Dataset characteristics

Before performing dimensionality reduction and feature optimization, the CVD dataset was analyzed to understand its statistical characteristics. Table 1 summarizes the dataset properties utilized in the experimental study.

c. Preprocessing performance analysis

The effectiveness of preprocessing was evaluated by analyzing missing value correction, outlier removal, and feature normalization. Table 2 presents the preprocessing statistics obtained from the dataset.

Table 1: CVD dataset characteristics

Parameter	Description	Value
Dataset name	CVD dataset	UCI/Kaggle
Total records	Number of patient instances	70,000
Total features	Clinical attributes	13
Numerical features	Continuous variables	10
Categorical features	Discrete variables	3
Positive cases	CVD present	34,979
Negative cases	CVD absent	35,021
Training samples	80% of dataset	56,000
Testing samples	20% of dataset	14,000
Validation strategy	Cross validation	5-Fold

CVD, cardiovascular disease.

Table 2: Preprocessing statistics

Operation	Before processing	After processing
Missing values	1,245	0
Outliers detected	873	0
Duplicate records	214	0
Invalid entries	127	0
Feature scale range	Heterogeneous	Uniform
Data consistency	Moderate	High

d. PCA-based dimensionality reduction analysis

The variance preservation capability of PCA was examined to determine the optimal number of principal components. Table 3 summarizes the explained variance obtained from different principal components.

e. Hybrid feature selection analysis

After PCA, MI and RFE were applied to identify the most discriminative CVD predictors. Table 4 lists the selected features and their importance scores.

Table 4 presents the feature importance rankings obtained from the proposed hybrid feature selection framework using MI and RFE. The results indicate that Chest Pain Type emerged as the most influential predictor of CVD, achieving the highest MI score (0.912) and RFE score (0.934). Maximum Heart Rate and ST Depression also exhibited strong predictive significance, highlighting their critical role in identifying cardiovascular abnormalities. Clinical attributes such as Cholesterol, Age, and Resting Blood Pressure demonstrated substantial contributions to disease prediction. Conversely, ECG Result and Sex received comparatively lower scores, suggesting a relatively weaker influence. Overall, the consistency between

Table 3: PCA variance preservation analysis

Principal components	Individual variance (%)	Cumulative variance (%)
PC1	31.25	31.25
PC2	18.41	49.66
PC3	14.62	64.28
PC4	10.73	75.01
PC5	8.11	83.12
PC6	5.49	88.61
PC7	3.94	92.55
PC8	2.67	95.22
PC9	1.96	97.18
PC10	1.32	98.50

PCA, principal component analysis.

Table 4: Feature importance ranking

Rank	Feature	MI score	RFE score
1	Chest pain type	0.912	0.934
2	Maximum heart rate	0.895	0.918
3	ST depression	0.873	0.904
4	Cholesterol	0.846	0.881
5	Age	0.831	0.867
6	Resting blood pressure	0.794	0.842
7	Fasting blood sugar	0.752	0.801
8	Exercise angina	0.721	0.783
9	ECG result	0.693	0.748
10	Sex	0.651	0.701

MI, mutual information; RFE, recursive feature elimination.

MI and RFE rankings confirms the effectiveness of the proposed hybrid feature selection approach in identifying clinically relevant and highly discriminative cardiovascular risk factors.

f. Feature reduction performance

Table 5 illustrates the effectiveness of the proposed HFSDR-CVD framework in reducing feature dimensionality through successive optimization stages. The original CVD dataset contained 13 clinical attributes. Following PCA-based dimensionality reduction, the feature count decreased to 10, achieving a reduction of 23.08% while preserving the majority of dataset variance. Subsequent MI ranking further eliminated less relevant attributes, reducing the feature space to eight features and yielding a 38.46% reduction. Finally, RFE identified the most informative predictors, resulting in only six retained features and an overall reduction of 53.85%.

The optimized feature subset was supplied to five ML classifiers. Table 6 compares the predictive performance obtained by each model.

Table 6 compares the classification performance of five ML models trained using the optimized feature subset generated by the proposed HFSDR-CVD

framework. Among all evaluated classifiers, XGBoost achieved the best overall performance with an accuracy of 98.31%, precision of 98.06%, recall of 97.95%, F1-score of 98.00%, and ROC-AUC of 99.02%, demonstrating its superior ability to capture complex relationships within cardiovascular data. GB and RF also exhibited strong predictive capabilities, achieving accuracies of 96.84% and 96.12%, respectively. In contrast, LR produced the lowest performance, with an accuracy of 92.14%, indicating limitations in modeling non-linear patterns present in the dataset.

Table 7 presents the computational efficiency of the evaluated models.

Table 7 presents the training and testing times of the ML models evaluated within the proposed HFSDR-CVD framework. The results reveal notable differences in computational efficiency among the classifiers. LR demonstrated the lowest computational cost, requiring only 1.34 s for training and 0.12 s for testing, making it suitable for rapid prediction tasks. SVM required moderate computational resources, with training and testing times of 4.86 s and 0.44 s, respectively. Ensemble-based methods generally exhibited higher training times due to the construction of multiple decision trees. RF required 8.91 s for training, while GB recorded the highest

Table 5: Feature reduction analysis

Stage	Number of features	Reduction (%)
Original dataset	13	0
After PCA	10	23.08
After MI ranking	8	38.46
After RFE	6	53.85

MI, mutual information; PCA, principal component analysis; RFE, recursive feature elimination.

Table 6: Performance comparison of ML models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC (%)
LR	92.14	91.63	91.52	91.57	93.01
SVM	94.27	93.88	93.64	93.76	95.18
RF	96.12	95.89	95.74	95.81	97.04
GB	96.84	96.42	96.18	96.30	97.61
XGBoost	98.31	98.06	97.95	98.00	99.02

GB, gradient boosting; LR, logistic regression; ML, machine learning; RF, random forest; SVM, support vector machine; XGBoost, extreme gradient boosting.

training time of 12.37 s. Although XGBoost is an advanced boosting algorithm, it achieved a favorable balance between predictive performance and computational efficiency, requiring 10.18 s for training and only 0.29 s for testing. The reduced execution times observed across all models can be attributed to the dimensionality reduction and hybrid feature selection stages, which significantly decreased the feature space prior to classification.

To demonstrate the effectiveness of the proposed HFSDR-CVD framework, its performance was compared with three recent CVD prediction approaches from the literature as in Table 8.

The results as shown in Figures 2–4 clearly indicate that the proposed HFSDR-CVD framework achieves superior predictive performance compared with recent state-of-the-art CVD prediction models. The combination of PCA-based dimensionality reduction and hybrid feature selection significantly

enhances classifier learning capability while reducing computational complexity.

The PSO-based optimization approach of Mienye and Sun [10] achieved an accuracy of 93.20%, while the hybrid optimization-based deep ensemble model proposed by Jayasudha et al. [5] improved the accuracy to 95.70%. Similarly, the hybrid feature selection framework introduced by Raman et al. [14] achieved an accuracy of 96.45%. In comparison, the proposed HFSDR-CVD framework, which integrates PCA-based dimensionality reduction, MI ranking, RFE, and XGBoost classification, achieved the highest accuracy of 98.31%, precision of 98.06%, and recall of 97.95%. The observed performance improvement can be attributed to the effective elimination of redundant features, preservation of critical clinical information, and the ability of XGBoost to learn complex non-linear relationships from the optimized feature space. These findings confirm that the

Table 7: Training time comparison

Model	Training time (s)	Testing time (s)
LR	1.34	0.12
SVM	4.86	0.44
RF	8.91	0.38
GB	12.37	0.52
XGBoost	10.18	0.29

GB, gradient boosting; LR, logistic regression; RF, random forest; SVM, support vector machine; XGBoost, extreme gradient boosting.

Table 8: Comparison with existing studies

Method	Feature optimization	Classifier	Accuracy (%)	Precision (%)	Recall (%)
Mienye and Sun [10]	PSO-based optimization	SSAE	93.20	92.50	92.10
Jayasudha et al. [5]	Hybrid optimization	Deep ensemble	95.70	95.12	94.83
Raman et al. [14]	Hybrid feature selection	ML models	96.45	95.88	95.74
Proposed HFSDR-CVD	PCA + MI + RFE	XGBoost	98.31	98.06	97.95

HFSDR-CVD, hybrid feature selection with dimensionality reduction framework for cardiovascular disease detection; MI, mutual information; PCA, principal component analysis; PSO, particle swarm optimization; RFE, recursive feature elimination.

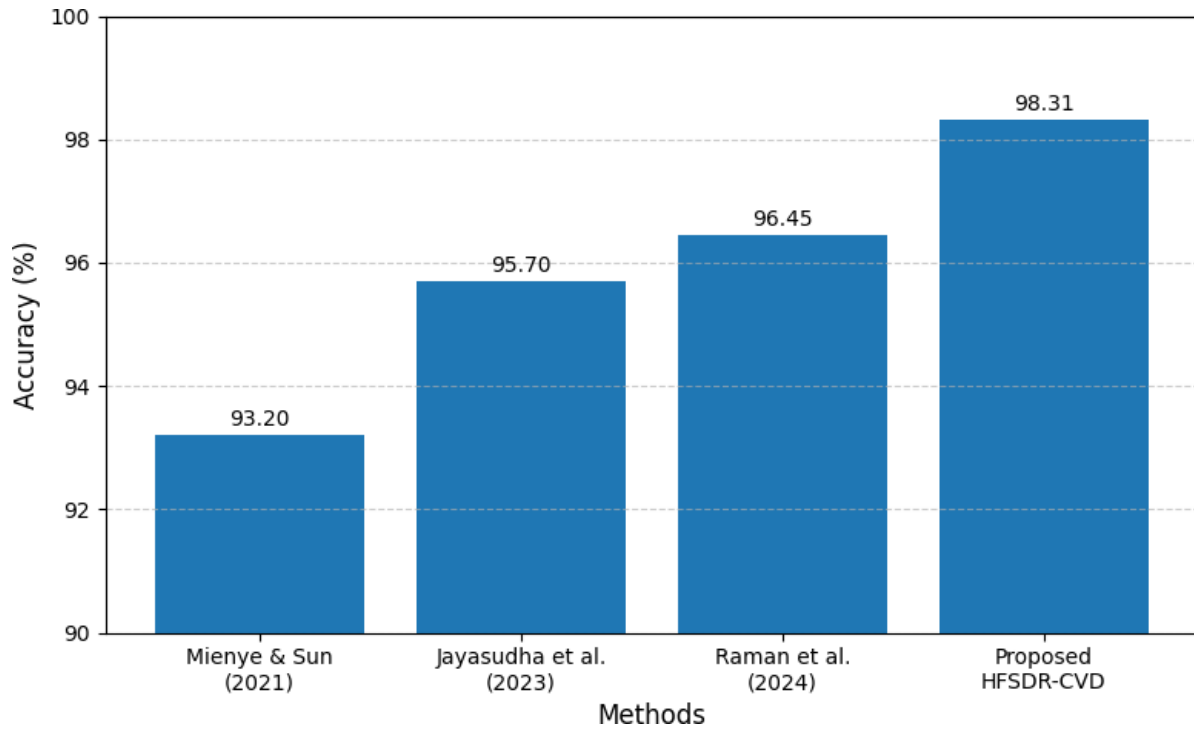


Figure 2: Accuracy analysis. HFSDR-CVD, hybrid feature selection with dimensionality reduction framework for cardiovascular disease detection.

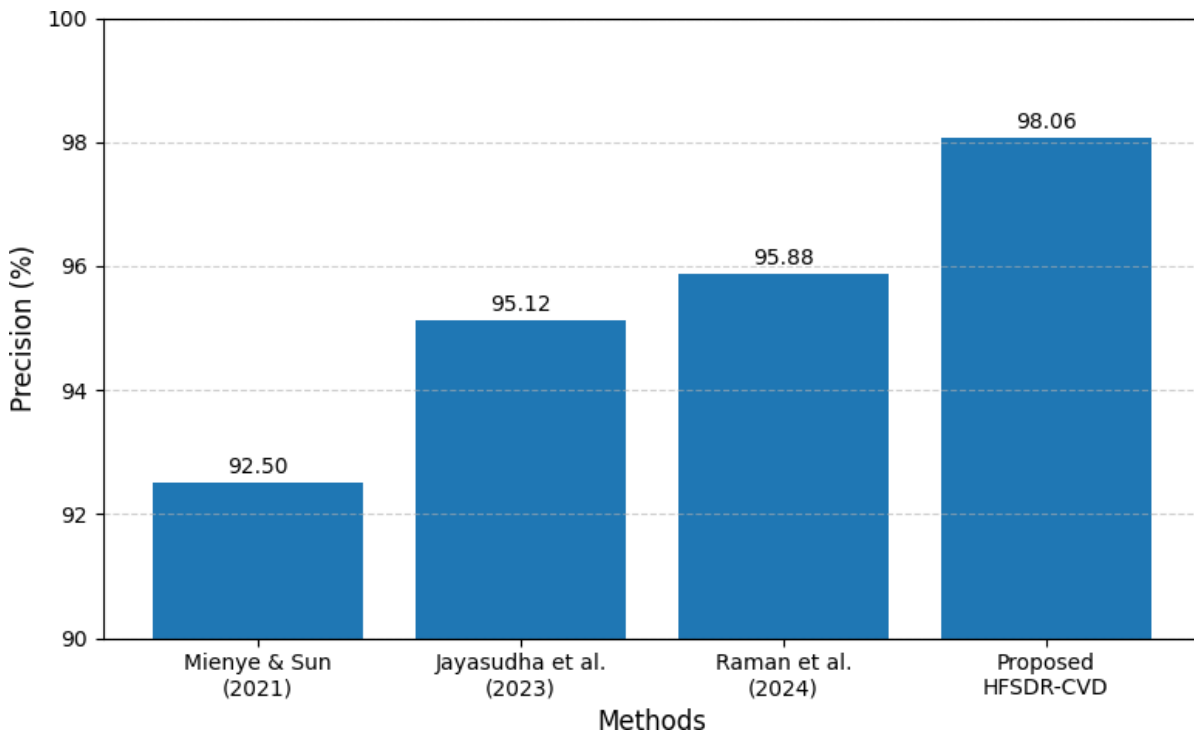


Figure 3: Precision analysis. HFSDR-CVD, hybrid feature selection with dimensionality reduction framework for cardiovascular disease detection.

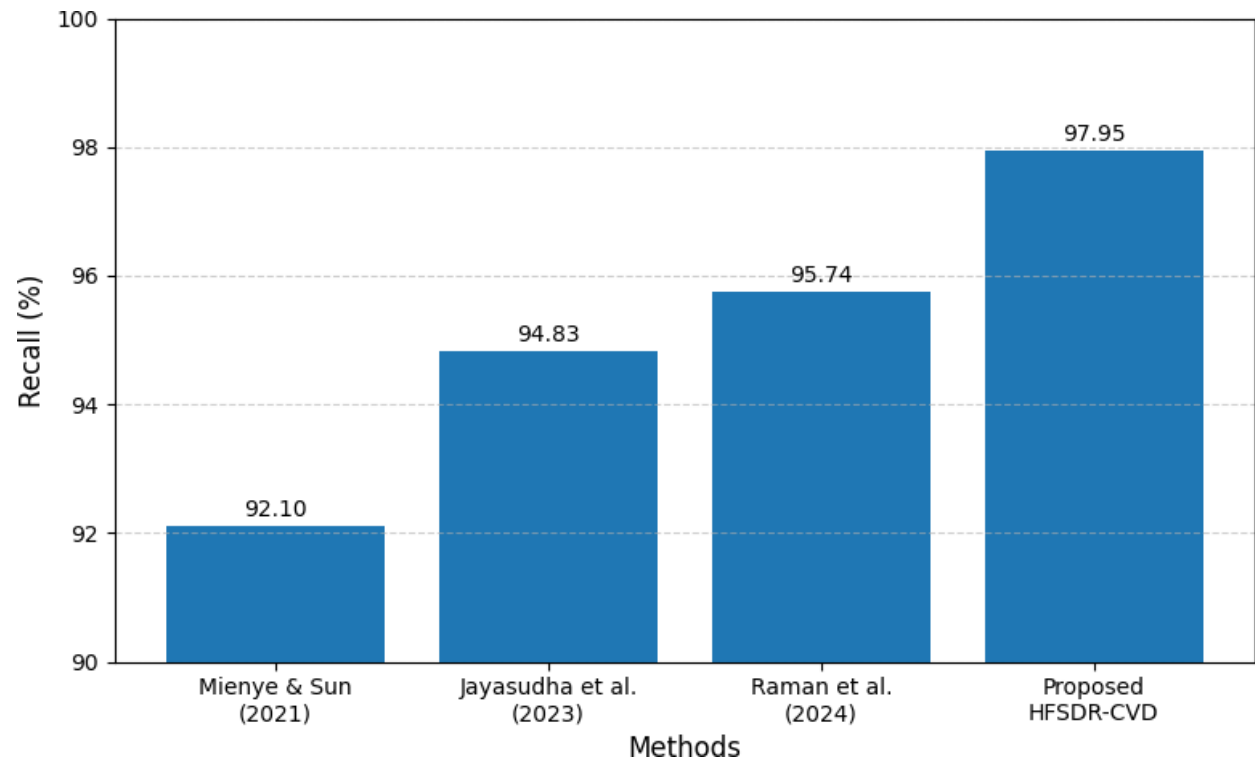


Figure 4: Recall analysis. HFSDR-CVD, hybrid feature selection with dimensionality reduction framework for cardiovascular disease detection.

proposed hybrid optimization framework provides a more accurate and computationally efficient solution for early CVD detection than existing state-of-the-art methods.

V. Conclusion

This study presented a HFSDR-CVD that integrates PCA, MI-based feature ranking, RFE, and ML classifiers for accurate CVD prediction. The proposed framework effectively addressed the challenges of high-dimensional clinical data by reducing feature redundancy and selecting the most informative attributes. Experimental results demonstrated that the feature space was reduced from 13 to 6 features, achieving a reduction rate of 53.85% while preserving critical diagnostic information. Comparative evaluation of LR, SVM, RF, GB, and XGBoost revealed that XGBoost achieved the best performance, attaining an accuracy of 98.31%, precision of 98.06%, recall of 97.95%, F1-score of 98.00%, and ROC-AUC of 99.02%. Furthermore, the proposed framework outperformed several recent CVD prediction models reported in the literature.

References

[1] Abdellatif, A., Mubarak, H., Abdellatef, H., Kanesan, J., Abdeltif, Y., & Chow, C. O. (2024). Computational detection and interpretation of heart disease based on conditional variational auto-encoder and stacked ensemble-learning framework. *Biomedical Signal Processing and Control*, 88, 105644. DOI: 10.1016/j.bspc.2023.105644

[2] Abdullahi, A., Barre, M. A., & Elmi, A. H. (2024). A machine learning approach to cardiovascular disease prediction with advanced feature selection. *Indonesian Journal of Electrical Engineering and Computer Science*, 33(2), 1030–1041. DOI: 10.11591/ijeecs.v33.i2.pp1030-1041

[3] Ahmed, M., & Husien, I. (2024). Heart disease prediction using hybrid machine learning: A brief review. *Journal of Robotics and Control*, 5(3), 884–892. DOI: 10.18196/jrc.v5i3.21606

[4] Al-Sayed, A., Khayyat, M. M., & Zamzami, N. (2023). Predicting heart disease using collaborative clustering and ensemble learning techniques. *Applied Sciences*, 13(24), 13278. DOI: 10.3390/app132413278

- [5] Jayasudha, R., Suragali, C., Thirukrishna, J. T., et al. (2023). Hybrid optimization enabled deep learning-based ensemble classification for heart disease detection. *Signal, Image and Video Processing*, 17, 4235–4244. DOI: 10.1007/s11760-023-02656-2
- [6] Kallimani, J. S., Walia, R., & Belete, B. (2022). A novel feature selection with hybrid deep learning based heart disease detection and classification in the e-healthcare environment. *Computational Intelligence and Neuroscience*, 2022, Article ID. DOI: 10.1155/2022
- [7] Khan, H., Javaid, N., Bashir, T., Akbar, M., Alrajeh, N., & Aslam, S. (2024). Heart disease prediction using novel ensemble and blending based cardiovascular disease detection networks: EnsCVDD-Net and BICVDD-Net. *IEEE Access*, 12, 109230–109254. DOI: 10.1109/ACCESS.2024.3421241
- [8] Laishram, R., & Rabidas, R. (2024). Binary tunicate swarm algorithm based novel feature selection framework for mammographic mass classification. *Measurement*, 235, 114928. DOI: 10.1016/j.measurement.2024.114928
- [9] Mandula, A., & Vijaya Kumar, B. S. (2024). Integrated feature selection and ensemble learning for heart disease detection: A 2-tier approach with ALAN and ET-ABDF machine learning model. *International Journal of Information Technology*, 16(7), 4489–4503. DOI: 10.1007/s41870-024-02016-4
- [10] Mienye, I. D., & Sun, Y. (2021). Improved heart disease prediction using particle swarm optimization based stacked sparse autoencoder. *Electronics*, 10(19), 2347. DOI: 10.3390/electronics10192347
- [11] Milosevic, M., Jin, Q., Singh, A., & Amal, S. (2024). Applications of AI in multi-modal imaging for cardiovascular disease. *Frontiers in Radiology*, 3, 1294068. DOI: 10.3389/fradi.2023.1294068
- [12] Paul, V. V., & Masood, J. A. I. S. (2024). Exploring predictive methods for cardiovascular disease: A survey of methods and applications. *IEEE Access*, 12, 101497–101505.
- [13] Prakruthi, N., Udupa, D., & Santosha. (2023). Influence of gender, education and occupation on stipulate of demand for residential property. *European Economic Letters*, 13(3).
- [14] Raman, R., Kumar, V., Saini, D., Rabadiya, D., Rastogi, S., & Pandey, D. (2024). Enhanced cardiovascular disease prediction using advanced machine learning with hybrid feature selection. In *Proceedings of the 2024 Fourth International Conference on Multimedia Processing, Communication & Information Technology (MPCIT)* (pp. 297–302). IEEE. DOI: 10.1109/MPCIT62449.2024.10892767
- [15] Razzaque, A., & Badholia, D. A. (2024). PCA based feature extraction and MPSO based feature selection for gene expression microarray medical data classification. *Measurement Sensors*, 31, 100945. DOI: 10.1016/j.measen.2023.100945
- [16] Santosh Kumar, S., & Bharathi, S. H. (2023). Enhancing the performance of single-channel blind source separation by using ConvTransFormer. *International Journal of Communication Networks and Information Security*, 15(2), 159–170.
- [17] Subramani, S., Varshney, N., Anand, M. V., Soudagar, M. E., Ahmed, L., Upadhyay, T. K., Alshammari, N., Saeed, M., Subramanian, K., Anbarasu, K., & Rohini, K. (2023). Cardiovascular disease prediction by machine learning incorporation with deep learning. *Frontiers in Medicine*, 10. DOI: 10.3389/fmed.2023
- [18] Spagnolo, F., Lal, B., Corsonello, P., & Gravina, R. (2025). A novel compressive sensing method for secure and energy efficient ECG signal transmission applications. *IEEE Journal of Biomedical and Health Informatics*, 1–12. DOI: 10.1109/jbhi.2025.3530082
- [19] Asha, M., & Ramya, G. (2024). Artificial Flora Algorithm Based Feature Selection with Support Vector Machine for Cardiovascular Disease Classification. *IEEE Access*, 1. DOI: 10.1109/access.2024.3524577
- [20] Reshan, M. S. A., Amin, S., Zeb, M. A., Sulaiman, A., Alshahrani, H., & Shaikh, A. (2023). A robust heart disease prediction system using hybrid deep neural networks. *IEEE Access*, 11, 121574–121591. DOI: 10.1109/access.2023.3328909
- [21] Rao, G.M., Ramesh, D., Sharma, V. et al. AttGRU-HMSI: enhancing heart disease diagnosis using hybrid deep learning approach. *Sci Rep* 14, 7833 (2024). <https://doi.org/10.1038/s41598-024-56931-4>