HYBRID FEATURE SELECTION STRATEGY FOR EFFICIENT HEART DISEASE CLASSIFICATION USING MACHINE LEARNING

D.Jeyasree[1], *A.Priyadharshini*[2]

Abstract

The fact that heart disease continues to be one of the top causes of death across the globe calls for the implementation of decision support systems that are driven by data to facilitate early detection. To find the most discrimination features from a structured datasets for heart disease, this study suggests a hybrid feature selection methodology that combines filter and wrapper approaches. Prior to employing wrapper-based refinement with recursive feature reduction and classifier feedback, the method uses statistical correlation and mutual information ranking to eliminate superfluous features. Several machine learning classifiers, such as logistic regression, support vector machine, gradient boosting, and random forest, are trained using the optimum feature subset. Compared to baseline models trained on the whole set of features, the hybrid feature selection technique considerably improves classification accuracy, decreases dimensional, and boosts interpretations, according to experimental evaluation. Early diagnosis of cardiac illness with enhanced accuracy and computational efficiency is made possible by the methodology, which lays a solid groundwork for clinical decision support.

Keywords: Heart disease, feature selection, filter methods, wrapper methods, machine learning, classification, healthcare analytic.

I. INTRODUCTION

Cardiovascular disease continues to represent a major global health challenge [1]. Timely prediction of heart disease can significantly improve patient outcomes and reduce healthcare costs. Clinical datasets often contain multiple heterogeneous attributes, including demographic, clinical, and physiological variables [2]. The presence of redundant or irrelevant features can degrade model performance and increase computational complexity [3].

Most traditional models use the full set of input attributes without systematically identifying the most predictive variables [4]. Such practices lead to over-fitting and interpretations challenges. A robust feature selection framework can enable more accurate predictions, lower training time, and greater clinical interpretations [5].

The core problem addressed in this work is to design an efficient hybrid feature selection methodology capable of identifying the minimal yet most informative feature subset for heart disease prediction.

- To develop a feature selection pipeline combining filter and wrapper strategies.
- To evaluate multiple machine learning classifiers trained on selected features.
- To assess the impact of feature selection on classification performance.
- To identify clinically relevant attributes contributing to heart disease prediction.

The proposed method reduces dimensional while retaining critical clinical indicators, enhancing both predictive accuracy and model interpretations. This contributes to more practical and explainable clinical decision support systems.

II. RELATED WORKS

Efficient heart disease classification using machine learning strongly depends on the adoption of effective feature selection strategies. These strategies play a critical role in enhancing predictive accuracy, reducing computational complexity, and improving model interpretations by identifying the most informative features within the datasets. Over the past decade, multiple studies have proposed and evaluated different feature selection approaches, each contributing unique insights into optimizing the predictive capabilities of classification algorithms for heart disease detection. A careful examination of the literature reveals that adaptive feature selection techniques, sensitivity analysis, ensemble and hybrid methods, and explainable machine learning approaches have significantly shaped the advancement of this field.

One prominent area of focus involves adaptive feature selection techniques, which have proven highly effective in

Department of Information Technology[1]
Karpagam Academy of Higher Education, Coimbatore, India[1]
harikavi24818 @gmail.com[1]

Department of Computer Applications[2]
Shri Nehru Maha Vidyalaya College of Arts and Science, Coimbatore[2]
phdpriyadharshini@gmail.com[2]

* Corresponding Author

refining datasets for improved classification performance. Mutual Information (MI) and Recursive Feature Elimination (RFE) have been particularly influential in this context. MI, which quantifies the dependency between input variables and target classes, allows for the identification of features with the highest predictive value, thereby enabling the elimination of irrelevant attributes before model training. Empirical evidence underscores its effectiveness: the combination of Support Vector Machine (SVM) with MI achieved a classification accuracy of 96.755% on the Cleveland Heart Disease datasets, while Random Forest integrated with MI achieved 97.4% on the Heart Statlog Cleveland datasets [6]. These findings indicate that filter-based feature selection strategies, when coupled with strong classifiers, can yield substantial gains in both accuracy and generalization. In a related study, RFE combined with Random Forest achieved 91% accuracy and an AUC of 92% on the Cleveland Clinic Heart Disease datasets, further validating the role of wrapper-based approaches in improving model performance [7].

Another significant line of investigation involves sensitivity analysis, particularly Variance-based Sensitivity Analysis (VSA), which evaluates how input variability influences predictive output. Unlike filter or wrapper methods that focus on direct relationships or classifier-driven selection, sensitivity analysis captures subtle interactions among variables, making it suitable for clinical datasets where feature interdependence often affect diagnostic outcomes. Saranya and Pravin (2021) demonstrated that VSA outperformed traditional wrapper techniques, reporting an accuracy of 87% and a sensitivity of 90.12% [8]. This evidence suggests that sensitivity-based approaches offer both performance gains and robustness, especially in medical datasets where clinical indicators often exhibit overlapping effects.

Recent studies have also highlighted the importance of ensemble and hybrid feature selection techniques in enhancing the predictive capability of machine learning models for heart disease. Ensemble strategies combine the predictive strengths of multiple classifiers, leading to more stable and generalization models. For example, a novel framework integrating X2 optimal feature selection with bagging and boosting classifiers achieved 97.84% and 98.44% accuracy, respectively [9]. These results emphasize the synergistic effect of combining robust feature selection with ensemble learners. Hybrid strategies have similarly shown promising outcomes. An approach that integrates Synthetic Minority Oversampling Technique (SMOTE) and Edited Nearest Neighbor (ENN) with RFE demonstrated substantial improvement in classification performance. When applied to heart disease datasets, XGBoost trained on the hybrid-selected features achieved an accuracy of 95.6%, outperforming baseline feature selection methods [10]. This underscores the value of combining filtering, oversampling, and wrapper strategies to address common issues such as class imbalance and feature redundancy.

The rise of explainable machine learning has further enriched the field of feature selection for heart disease prediction. Traditional models often function as black boxes, making it difficult to understand which features contribute most to the classification decision. Explainable techniques such as permutation importance and SHapley Additive explanations (SHAP) allow for feature impact analysis in a transparent manner. Aprianto and Anasanti (2025) employed these explainable techniques and found that resting electroencephalographic measurements, maximum heart rate, cholesterol level, and age were the most influential features for heart disease prediction [11]. By identifying these high-impact features, explainable methods not only enhance the transparency of machine learning models but also align predictive outputs with established clinical knowledge, supporting informed medical decision-making.

Another emerging trend involves advanced ensemble learning techniques, particularly stacking classifiers, which have been shown to deliver superior results compared to traditional single-model approaches. A recent study employed a stacking model that combined Boosted Decision Trees, Extra Trees, and LightGBM to achieve perfect classification accuracy across different feature selection techniques [12]. This result illustrates the potential of integrating sophisticated feature selection methods with ensemble models capable of capturing complex nonlinear interactions in medical data. Such hybridized approaches not only enhance predictive accuracy but also ensure that the model remains robust across different feature configurations.

In [13] researches emphasized that the critical role of ensemble-based feature selection and classification techniques in improving heart disease prediction accuracy. The work integrates multiple ensemble learners to enhance model stability and generalization, resulting in robust predictive performance [13].

In [14] researchers presented a comparative evaluation of classical machine learning algorithms, including Logistic Regression, Random Forest, SVM, and KNN, applied to the UCI Heart Disease datasets. The study highlights the significance of model selection and evaluation metrics in achieving reliable diagnostic outcomes.

In [15] authors focused on optimized feature selection strategies to improve coronary artery disease prediction. The research demonstrates that careful selection of features coupled with comparative model analysis yields enhanced prediction accuracy and clinical applicability.

These studies collectively demonstrate that no single feature selection strategy can universally outperform others across all contexts. The choice of method depends on multiple factors, including datasets characteristics, class distribution, feature dimensional, and desired trade-offs between accuracy and interpretations. Filter-based techniques such as MI offer computational efficiency and high-ranking stability, while wrapper methods like RFE provide more refined selection through classifier feedback. Sensitivity analysis adds an additional layer of robustness by accounting for inter-feature variability. Hybrid approaches that combine oversampling and feature selection methods address real-world challenges like class imbalance. Furthermore, the integration of explainable machine learning enhances the interpretations of predictive models, which is especially critical in medical domains where clinical trust and regulatory compliance are paramount. Finally, stacking and other advanced ensemble learning strategies show that combining well-selected features with powerful learners can produce near-optimal performance.

Related research highlights the evolution from simple filter methods to complex hybrid and explainable approaches for heart disease classification. As datasets grow richer and models become more sophisticated, the role of effective and interpretative feature selection will remain central to building high-performing and clinically trustworthy diagnostic systems.

III. FEATURE SELECTION METHODOLOGIES

The workflow illustrated in Figure 1 outlines a structured pipeline for data-driven model development. The process begins with data acquisition, followed by re-processing steps such as missing value handling, encoding, and scaling to ensure data quality. A hybrid feature selection strategy integrates filter methods, embedded methods, and wrapper methods to identify the most informative attributes. Imbalance handling techniques such as SMOTE are then applied to address class distribution issues. The refined datasets is used for classification through algorithms such as Logistic Regression, SVM, and Random Forest, with subsequent performance evaluation to assess model effectiveness (Figure 1).

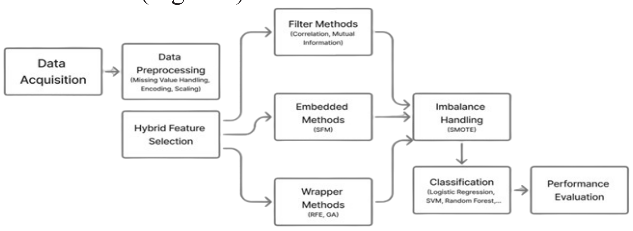
Data Acquisition
Data Preprocessing (Missing Value Handling, Encoding, Scaling)
Hybrid Feature Selection
Filter Methods (Correlation, Mutual Information)
Embedded Methods (SFM)
Wrapper Methods (RFE, GA)
Imbalance Handling (SMOTE)
Classification (Logistic Regression, SVM, Random Forest,...
Performance Evaluation

Figure 1. Methodology Diagram.

A. Data Preprocessing

Data pre-processing plays a crucial role in ensuring the quality, consistency, and reliability of input features before applying any feature selection or classification algorithm. In this study, the pre-processing pipeline was designed to minimize noise, handle inconsistencies, and improve the numerical stability of the learning models. The first step involved handling missing values to ensure data completeness. Records with excessive missing attributes were removed, while for remaining instances, missing numerical values were imputed using mean or median values to preserve the overall data distribution. This process helped prevent bias introduced by incomplete observations, ensuring that the statistical structure of the datasets remained intact.

The next step focused on encoding categorical attributes into numerical representations to make the datasets suitable for machine learning algorithms. Since most classification algorithms require numerical input, categorical variables such as gender, chest pain type, and exercise-induced angina were transformed using one-hot encoding or label encoding. For example, a categorical variable with k unique categories was converted into a binary matrix of dimension X_c with k unique categories was converted into a binary matrix of dimension $n \times k$ where each column represents a distinct category. This transformation can be mathematically represented as $X_{\text{encoded}} = \text{OneHot}(X_c) \in R^{n \times k}$ here, n is the sample count and k is the category count.

Normalization and standardization were applied to continuous variables to maintain uniform scaling across features. Features such as cholesterol level, resting blood pressure, and maximum heart rate exhibited different ranges, which could bias the learning process if left unmeasured. Min–max normalization was used to re-scale these attributes to a fixed range, typically ([0, 1]), using the transformation as in equation 1:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

is the original value, and x_{min}d x_{max}e the minimum and maximum values of the feature, respectively. This change made sure each feature played an equal role in model training.

In addition to normalization, standardization was employed to center the data around zero and adjust it to unit variance, improving the convergence behavior of gradient-based optimization algorithms. This process was defined as in equation 2:

$$z = \frac{x - \mu}{\sigma} \tag{2}$$

here μ is the mean and σ is the feature's standard deviation. Standardization improves numerical stability, especially for algorithms sensitive to feature scales like SVM and gradient methods.Through this structured preprocessing pipeline comprising missing value treatment, categorical encoding, normalization, and standardization the datasets was transformed into a consistent and well-conditioned form, allowing the subsequent feature selection and classification stages to operate with improved stability and accuracy. This ensures that the learning process is driven by genuine data patterns rather than scale disparities or incomplete information.

B. Filter-based Selection

Filter-based feature selection serves as an essential stage for reducing dimensional by identifying the most informative and least redundant features before applying more complex selection or classification methods. This study employed correlation analysis and mutual information (MI) ranking as primary filter techniques. These methods evaluate features based on statistical properties and their relationship with the target class, independent of any specific machine learning algorithm.

The first technique, correlation analysis, aims to remove redundant or highly collinear features. If two features are strongly correlated, one can be eliminated without significantly affecting predictive performance. The Pearson correlation coefficient served as the metric to assess the linear association among features, and is defined as in equation 3:

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (3)$$

where $\rho_{X,Y}$ denotes the correlation coefficient between features *X* and *Y, cov(X,Y)* is their covariance, and σ_X, σ_Y represent standard deviations of X and Y respectively. The coefficient ranges between (-1) and (+1). A value near +1 shows strong positive correlation, near -1 strong negative, and around 0 little or no linear relationship. In this work, a predefined threshold (e.g., $(|\rho| > 0.8)$) was used to identify and eliminate one of the highly correlated feature pairs to avoid multicollinearity and over fitting.

The second method was mutual information ranking, which measures how much one variable tells us about another. Unlike correlation, MI detects both linear and non-linear ties, making it well-suited for clinical data. Mutual information between feature X and target Y is defined the below equation 4:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x),p(y)} \quad (4)$$

The notation $(\mathrm{p}(x,y))$ refers to the joint probability for variables (X) and (Y), whereas $(\mathrm{p}(x))$ and $(\mathrm{p}(y))$ indicate the individual marginal probabilities of (X) and (Y) respectively. A larger I(X;Y) means the feature is more strongly linked to the target and holds greater predictive value.

In practice, for each feature, MI scores were computed and ranked in descending order. The top-ranked features were retained for further modeling, while low-ranking features were discarded. This step ensures that only the most informative features with high predictive relevance are passed on to subsequent stages of the pipeline.

By integrating correlation analysis to eliminate redundant attributes and mutual information ranking to select the most informative ones, the dimensional of the datasets were reduced effectively. This combination improves computational efficiency, enhances model generalization, and establishes a solid foundation for wrapper-based refinement in the subsequent steps.

C. Wrapper-based Refinement

Wrapper-based feature selection optimizes feature subsets by testing their effect on model performance. Unlike filter methods, wrappers train algorithms on multiple feature combinations, choosing the best set for prediction. In this study, RFE with cross-validation was used to further refine features identified in the filter stage. Classifier feedback was used as a performance signal to guide the elimination of less informative features and retain only the most relevant ones.

RFE involves training a model, ranking features by their importance, and recursively removing the least important one until only the desired number remains. Models like SVM, LR, and RF provide feature importance scores that reflect each feature's impact on prediction.

Let $D = (x_i, y_i)_{i=1}^{n}$ denote the training datasets with n instances, where $x_i \in R^d$ represents a feature vector of dimension d, and y_i is the target label. For a linear model such as logistic regression or SVM, the importance of feature j can be represented by the absolute value of its learned weight: $I_j = |w_j|$, where w_j denotes the weight assigned to feature j after fitting the model. Features with lower are considered less important and are removed in subsequent iterations. Tree-based classifiers like RF often measure importance using Gini importance or mean decrease in impurity in equation 5:

$$I_j = \sum_{t \in T_j} p(t)\,, \Delta i(t) \quad (5)$$

where T_j is the set of nodes where feature j is used for splitting, p(t) is the proportion of samples reaching node t, and $\Delta i(t)$ is the decrease in impurity produced by that split. This score reflects how much a feature contributes to improving node homogeneity in the decision tree.

During each iteration of recursive feature elimination (RFE), the feature demonstrating the lowest importance score is excluded, after which the model is retrained using the remaining features and its performance is assessed. This procedure is repeated until the desired subset of features is obtained. To enhance the robustness and generalization of the selection process, cross-validation has been incorporated into the RFE methodology. In a k-fold cross-validation approach, the datasets is partitioned into k equal segments, with the model being trained on k-1 folds and validated against the remaining fold. The overall performance is averaged across folds: $\text{CV\_Score} = \frac{1}{}\sum_{i=1}^{k} \text{Accuracy}_i$ where Accuracy_i represents the model accuracy on the i-th validation fold. This score provides reliable feedback to determine whether eliminating a particular feature improves or degrades model performance.

The RFE process can be formally expressed as in equation 6:

1. Train the base classifier on the current feature set F.
2. Compute the importance score I_j for each feature $f_j \in F$.
3. Rank features in ascending order of importance.
4. Eliminate the least important feature $f_{\min}$.
5. Repeat until the desired number of features remains.

$$F_{t+1} = F_t \arg\min_{f_j \in F_t} I_j \tag{6}$$

This iterative process ensures that only the most relevant features are retained, while redundant or weakly contributing features are systematically removed. By combining classifier feedback with cross-validation performance, this wrapper-based refinement improves both predictive accuracy and model stability.

D. Classifier Integration

After the feature selection stages, the most informative and non-redundant subset of attributes was used to train multiple machine learning classifiers. This step ensures that the refined feature set enhances model performance, generalization, and interpretations. In this study, LR, RF, SVM, and Gradient Boosting classifiers were employed. Additionally, grid search-based hyper-parameter tuning with cross-validation was used to identify optimal model configurations.

Logistic Regression

LR is a straightforward linear classifier often used in medical prediction because it's easy to interpret. It calculates the probability of an instance being in the positive class via the logistic function given in equation 7:

$$P(y = 1 \mid X) = \frac{1}{1+e^{-(w^T x+b)}} \tag{7}$$

Here, x denotes the input feature vector, indicates the corresponding learned weight vector, and b represents the bias term. The model parameters are determined by maximizing the log-likelihood function in equation 8:

$$\mathcal{L}(w,b) = \sum_{i=1}^{n} [y_i \log P_i + (1 - y_i) \log(1 - P_i)] \tag{8}$$

where P_i is the predicted probability for sample i, and y_i is the corresponding ground truth label. Logistic Regression serves as a strong baseline for evaluating the discriminating power of selected features.

Support Vector Machine

An SVM classifier constructs a hyperplane aimed at optimally dividing different classes within the feature space. When dealing with data that can be separated by a straight line, this separating margin is described as follows: $f(x) = w^T x + b = 0.$ The optimization problem aims to maximize the margin $\frac{2}{|w|}$ subject to: $y_i(w^T x_i + b) \geq 1 \forall i$ where $y_i \in -1, +1$ denotes class labels. For non-linear problems, kernel functions such as the radial basis function (RBF) can be applied to map data into a higher-dimensional space.

Random Forest

RF is an ensemble method that improves classification and reduces over fitting by combining several decision trees, each trained on a bootstrap sample. Predictions are made through majority voting. The prediction function for Random Forest can be expressed as $\hat{y} = \text{mode} h_t(x)_{t=1}^{T}$ where h_t refers to tree t's prediction, with T as the total number of trees. Feature importance in RF is measured by the mean decrease in impurity: $I_j = \sum_{t=1}^{T} \sum_{n \in N_{j,t}} p(n), \Delta i(n)$ where $N_{j,t}$ represents the set of nodes where feature *j* is used in tree *t,p(n)* is the proportion of samples reaching node *n,* and $\Delta i(n)$ is the reduction in impurity at that node.

Gradient Boosting

Gradient Boosting creates an ensemble of weak learners usually decision trees sequentially, with each model refining the errors of its predecessor. The prediction function is given in below equation 9:

$$F_M(x) = \sum_{m=1}^{M} \gamma_m h_m(x) \tag{9}$$

where h_m is the weak learner at iteration m, γ_m is the corresponding weight, and M is the total number of boosting stages. Each learner minimizes the negative gradient of a differentiable loss function L, typically logistic loss for binary classification: $\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\big(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\big).$ This iterative fitting of residuals allows the model to learn complex decision boundaries with high accuracy.

Hyperparameter Tuning

Hyperparameters significantly influence model performance. To identify optimal settings, grid search was employed with k-fold cross-validation. For each combination of hyper-parameters θ, the model's performance was evaluated as: $\text{CV\_Score}(\theta) = \frac{1}{k}\sum_{i=1}^{k} \text{Accuracy}_i(\theta)$ where $\text{Accuracy}_i(\theta)$ is the accuracy obtained on the i-th validation fold. The best hyper-parameter configuration was selected based on the maximum cross-validation score.

This classifier integration stage ensures that the final predictive models are trained on an optimized feature subset and fine-tuned hyper-parameters, leading to improved generalization and robust performance in heart disease classification.

IV. RESULTS AND DISCUSSION

A. Experimental Setup

The methodology was evaluated on a structured heart disease datasets [16]. The experiments were executed in Python using scikit-learn. Performance was compared between models trained with all features and those trained with the selected subset.

B. Feature Selection Outcome

- The original datasets contained 13 features.
- The hybrid method selected 7 key attributes.

Feature importance analysis indicated that variables such as "cholesterol level," "maximum heart rate," and "exercise induced angina" contributed strongly to prediction.

C. Classification Performance

The comparative evaluation of multiple feature selection strategies under both standard (imbalanced) and imbalance-aware (SMOTE) classification settings provides critical insight into how feature selection interacts with class distribution in structured tabular health data. When feature selection was applied without any re-balancing, all methods demonstrated high overall accuracy (≈0.91) but simultaneously exhibited very low recall (≈0.05–0.10) and consequently poor F1-scores (≈0.08–0.16). This discrepancy arises from the classifier's tendency to favor the majority class, achieving high accuracy through correct classification of negative samples while failing to capture positive instances representing the minority class. In such imbalanced scenarios, traditional metrics like accuracy become misleadingly optimistic, masking poor sensitivity. Notably, even relatively advanced feature selectors like logistic regression–based embedded selection and genetic algorithms failed to mitigate this issue when class imbalance remained unaddressed, suggesting that feature selection alone is insufficient to handle skewed distributions.

When SMOTE-based oversampling was introduced after feature selection, the results changed substantially. Across all methods, recall increased dramatically to the 0.69–0.77 range, indicating a much stronger ability of the classifiers to correctly identify minority class cases. F1-scores improved correspondingly to 0.30–0.34, despite a reduction in accuracy to approximately 0.69–0.75. This accuracy drop reflects a shift in the classifier's decision boundary toward greater sensitivity rather than a true degradation in predictive quality. ROC-AUC scores remained broadly stable, confirming that re-balancing altered the operating point on the ROC curve without degrading ranking ability. This outcome demonstrates the crucial role of integrating imbalance handling techniques alongside feature selection to obtain reliable and clinically meaningful results.

Among the methods evaluated, the logistic regression–based SelectFromModel (SFM_LR) method emerged as the top performer after SMOTE, achieving the highest F1-score (0.3400) and maintaining strong ROC-AUC (0.8243). Its strength lies in the fact that logistic regression directly aligns with the downstream classifier, selecting features most relevant for linear decision boundaries. This alignment minimizes distributional shift between feature selection and classification phases. However, the main limitation of SFM_LR is its model dependence its effectiveness is tied to linear assumptions, which may not generalize optimally if more complex classifiers are applied later. In contrast, RFE_LR yielded slightly lower but very close performance, indicating that recursive elimination, though more computationally expensive, does not yield substantial gains over embedded selection in this datasets.

Filter-based methods such as ANOVA_F, Mutual_Info, and Chi2 exhibited similar trends. Before re-balancing, these filters performed poorly in recall, as they are purely univariate and do not address class imbalance. After SMOTE, their recall values rose substantially, but precision declined more than in embedded methods. Chi2 produced the highest recall (0.7729) among all filters, suggesting that its selected features had strong discriminatory power for the minority class. Yet, its lower precision compared to SFM_LR reflects a weaker control over false positives, likely due to its purely statistical (rather than model-aligned) nature. The weakness of filter methods in imbalanced settings is their independence from classifier behavior, which can lead to sub-optimal feature subsets if not accompanied by downstream calibration.

The Correlation and Variance Threshold methods, which simply prune redundant or low-variance features, showed surprisingly strong performance after SMOTE, achieving F1-scores close to the top methods. Their strength lies in simplicity and computational efficiency: these methods preserve a broad but less redundant set of features, allowing the classifier to leverage more information after re-balancing. However, their weakness is the lack of direct class-discrimination focus, which can lead to inclusion of irrelevant or weakly predictive features in more complex datasets. Nonetheless, their competitive performance in this study illustrates that, when coupled with class re-balancing, even simple feature selection can be remarkably effective.

Tree-based embedded methods such as RFE_RF, SFM_RF, RF_Top10, and SFM_DT showed consistent recall improvements after SMOTE, like other methods, but achieved slightly lower F1-scores overall. Their ROC-AUC scores remained stable, indicating decent ranking performance. These methods capture nonlinear relationships but select features based on importance measures that may not align perfectly with logistic regression, the downstream classifier used in evaluation. This model mismatch likely explains their lower precision and F1 relative to linear selectors. Their strength lies in their ability to discover nonlinear predictive structures, but their weakness in this pipeline is lack of compatibility with the final classification model, which reduces their practical benefit in this specific setup.

Dimensional reduction via PCA performed best among non-rebalanced runs, with an F1-score of 0.1603, reflecting its ability to compress information into compact components even under imbalance. However, its post-SMOTE performance (F1=0.3328) remained behind supervised selectors. This difference stems from PCA's unsupervised nature: it preserves variance without prioritizing class discrimination, which can dilute the signal from minority class samples. PCA's strength lies in dimensional reduction and noise suppression, but its weakness is the lack of class-awareness, making it less effective for minority detection tasks.

Taken together, these findings demonstrate that imbalance handling is the decisive factor in transforming feature selection from a purely technical preprocessing step into a clinically meaningful component of predictive modeling. Feature selection methods alone cannot overcome skewed class distributions; their effectiveness becomes fully realized only when combined with re-balancing strategies. Linear embedded selectors like SFM_LR excel under this framework because of their direct alignment with the logistic decision boundary, while filter and tree-based methods display trade-offs between simplicity, nonlinear capacity, and precision. This critical comparison underlines the need for joint optimization of feature selection and class re-balancing, rather than treating them as isolated steps. Such integration ensures both model interpretations and predictive reliability, particularly in sensitive domains like healthcare analytic where minority class detection is often the central objective.

Table 1. Performance of Feature Selection Methods Without Imbalance Handling

Method	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
PCA	0.912725	0.504735	0.095315	0.160349	0.805834
ANOVA_F	0.912944	0.512422	0.088519	0.150961	0.822429
GA	0.913147	0.519744	0.087089	0.149181	0.823605
SFM_LR	0.913257	0.523810	0.086552	0.148557	0.823978
Variance_Threshold	0.912819	0.508439	0.086195	0.147401	0.826232
Correlation	0.912819	0.508439	0.086195	0.147401	0.826232
RFE_LR	0.913085	0.517838	0.085658	0.147000	0.824455
Mutual_Info	0.912522	0.498301	0.078684	0.135907	0.820102
Chi2	0.912600	0.501229	0.072961	0.127381	0.814893
RFE_RF	0.912178	0.481481	0.058119	0.103718	0.807561
SFM_RF	0.912381	0.490536	0.055615	0.099904	0.804644
RF_Top10	0.912209	0.481541	0.053648	0.096541	0.805059
SFM_DT	0.912725	0.509881	0.046137	0.084618	0.798974

As illustrated in figure 2, all feature selection methods achieve high accuracy but exhibit poor recall and F1-scores in the imbalanced setting, confirming the classifier's bias toward the majority class.

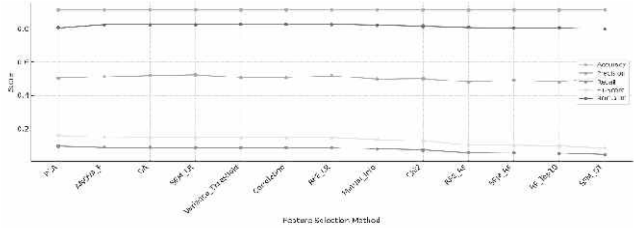

Figure 2. Performance analysis without SMOTE

Table 2. Performance of Feature Selection Methods With SMOTE (Imbalance Handling)

Method	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
PCA	0.757360	0.219052	0.692060	0.332774	0.806711
ANOVA_F	0.737707	0.216257	0.762160	0.336917	0.822666
GA	0.738739	0.216840	0.761266	0.337536	0.822799
SFM_LR	0.741506	0.218870	0.761624	0.340026	0.824248
Variance Threshold	0.736847	0.217116	0.771280	0.338846	0.825690
Correlation	0.736847	0.217116	0.771280	0.338846	0.825690
RFE_LR	0.737973	0.217134	0.766452	0.338400	0.824610
Mutual_Info	0.734095	0.214239	0.765200	0.334755	0.820404
Chi2	0.717178	0.204437	0.772890	0.323346	0.814813
RFE_RF	0.711002	0.198475	0.758763	0.314646	0.807273
SFM_RF	0.704248	0.195410	0.764306	0.311244	0.804543
RF_Top10	0.706140	0.196441	0.763948	0.312521	0.804835
SFM_DT	0.693241	0.189233	0.763770	0.303317	0.798638

In contrast, figure 3 clearly shows that applying SMOTE increases recall and F1-scores substantially while maintaining stable ROC-AUC values, reflecting a more balanced and effective decision boundary.

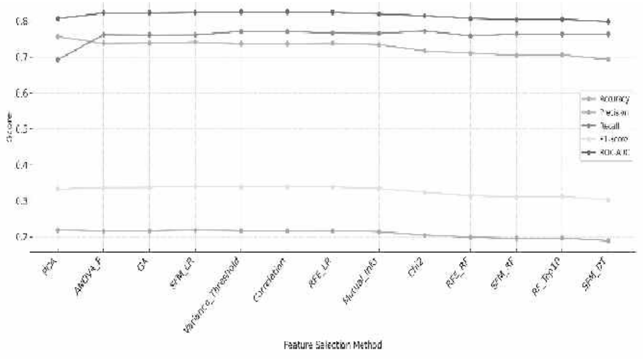

Figure 3. Performance analysis with SMOTE

V. CONCLUSION

The experimental results demonstrate that feature selection alone is insufficient to address class imbalance in heart disease classification. Without SMOTE, all methods achieved high accuracy (≈0.91) but exhibited poor recall (0.05–0.10) and low F1-scores (0.08–0.16), indicating a strong bias toward the majority class. After applying SMOTE, recall increased substantially to the range of 0.69–0.77 and F1-scores improved to 0.30–0.34, while accuracy decreased to 0.69–0.75 due to a more balanced decision boundary. The SFM_LR method achieved the best performance with F1 = 0.3400, recall = 0.7616, and ROC-AUC = 0.8242. Correlation and Variance Threshold methods also showed strong performance (F1 = 0.3388), indicating that even simple selectors benefit from re-balancing. Tree-based methods maintained stable ROC-AUC but slightly lower F1-scores. Combining SMOTE with appropriate feature selection significantly enhances minority class detection and produces more reliable classification performance for heart disease diagnosis.

REFERENCES

[1] Kasartzian, D. I., & Tsiampalis, T. (2025). Transforming cardiovascular risk prediction: a review of machine learning and artificial intelligence innovations. Life, 15(1), 94.

[2] Maruotto, I., Ciliberti, F. K., Gargiulo, P., & Recenti, M. (2025). Feature Selection in Healthcare Datasets: Towards a Generalizable Solution. Computers in Biology and Medicine, 196, 110812.

[3] Karimi, M., Karimi, Z., Khosravi, M., Delaram, Z., Dehsheikhim, M. H., Najafabadi, S. A., ... & Tavakoli, N. (2025). Feature Selection Methods in Big Medical Databases: A Comprehensive Survey. International Journal of Theoretical & Applied Computational Intelligence, 181-209.

[4] Ahmed, K. R., Ansari, M. E., Ahsan, M. N., Rohan, A., Uddin, M. B., & Rivin, M. A. H. (2025). Deep learning framework for interpretable supply chain forecasting using SOM ANN and SHAP: KR Ahmed et al. Scientific Reports, 15(1), 26355.

[5] Wang, H., Zhang, M., Mai, L., Li, X., Bellou, A., & Wu, L. (2025). An effective multi-step feature selection framework for clinical outcome prediction using electronic medical records. BMC Medical Informatics and Decision Making, 25(1), 1-15.

[6] Oleiwi, Z., AlShemmary, E. N., & Al-augby, S. (2023). Adaptive Features Selection Technique for Efficient Heart Disease Prediction. https://doi.org/10.29304/jqcm.2023.15.1.1137

[7] Maulana, A., Faisyal, F. R., Tarmizi, F. K., Abidin, T. F., & Riza, H. (2023). Optimizing Heart Disease Classification: Exploring the Impact of Feature Selection and Performance of Machine Learning Algorithms. https://doi.org/10.1007/978-981-99-7969-1_20

[8] Saranya, G., & Pravin, A. (2021). An Efficient Feature Selection Approach using Sensitivity Analysis for Machine Learning based Heart Desease Classification. International Conference on Communication Systems and Network Technologies. https://doi.org/10.1109/CSNT51715.2021.9509673

[9] Jaisinghani, K. S., & Malik, S. (2023). Enhanced Feature Selection and Extraction for Ensemble Machine Learning-based Classification of Heart Disease based on ECG. https://doi.org/10.1109/i-smac58438.2023.10290337

[10] Effective Feature Selection for Improved Prediction of Heart Disease. (2022). Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-030-93314-2_6

[11] Aprianto, K., & Anasanti, M. D. (2025). Classifying Heart Disease through Fusion of Multi-Source Datasets: Integration of Feature Selection and Explainable Machine Learning Techniques. IJCCS. https://doi.org/10.22146/ijccs.92395

[12] Mrs.Deepa, Mrs. D., & Krishna, M. H. (2025). Predicting Heart Diseases Using Machine Learning and Different Data Classification Techniques. International Journal of Engineering and Science Invention. https://doi.org/10.35629/6734-14081424

[13] Sreekumari, S., Bhalla, R., & Singh, G. (2025). Feature Selection and Model Evaluation for Heart Disease Prediction Using Ensemble Methods. Procedia Computer Science, 259, 1282-1295.

[14] Nasution, N., Hasan, M. A., & Nasution, F. B. (2025). Predicting Heart Disease Using Machine Learning: An Evaluation of Logistic Regression, Random Forest, SVM, and KNN Models on the UCI Heart Disease Dataset. IT Journal Research and Development, 9(2), 140-150.

[15] Olawade, D. B., Soladoye, A. A., Omodunbi, B. A., Aderinto, N., & Adeyanju, I. A. (2025). Comparative analysis of machine learning models for coronary artery disease prediction with optimized feature selection. International Journal of Cardiology, 133443.

[16] Pytlak, K. (2022). Personal Key Indicators of Heart Disease [Data set]. Kaggle. https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease/data