Hybrid Feature Selection and Ensemble Learning for Multi-Disease Prediction in Smart Healthcare Systems

Dr. Deepali Atul Godse[1*], Dr. Suwarna Gothane[2], Vijaya Balpande[3], Dr. Manoj Vasantrao Bramhe[4], Miss. Kalyani P. Karule[5], Amruta Tejaskumar Mokashi[6]

[1]Professor, Computer Engineering, Bharati Vidyapeeth's College of Engineering for Women, Pune, India
deepali.godse@bharatividyapeeth.edu
[2]Artificial Intelligence and Data Science
DYPCOE, Akurdi, Pune - 411044 Maharashtra, India
gothane.suvarna@gmail.com
[3]Department of Computer Science and Engineering,
Priyadarshini College of Engineering, Nagpur, India
vpbalpande15@gmail.com
[4]Professor, Department of Computer Science and Engineering
St. Vincent Pallotti College of Engineering and Technology, Nagpur, India
mbramhe@stvincentngp.edu.in
[5]Assistant Professor, Department of Computer Technology
Yeshwantrao Chavan College of Engineering, Nagpur, India
kkarule92@gmail.com
[6]Department of Chemical Engineering
Vishwakarma Institute of Technology, Pune, Maharashtra, 411037, India
amruta.mokashi2@vit.edu

HOW TO CITE:
Deepali Atul Godse, Suwarna Gothane, Vijaya Balpande, Manoj Vasantrao Bramhe, Kalyani P. Karule, Amruta Tejaskumar Mokashi (2026). Hybrid Feature Selection and Ensemble Learning for Multi-Disease Prediction in Smart Healthcare Systems. International Journal of Special Education, 41(7s), 150-163.

ABSTRACT:
Smart healthcare systems with high dimensional clinical data, feature redundancy and limited generalization of conventional models pose a challenge to the accurate prediction of multiple diseases. The main goal of this study is to overcome these problems by introducing a hybrid feature selection and ensemble learning framework which improves the prediction accuracy and also has better computational efficiency. The method is based on two-stage feature optimization process, which combines ReliefF to rank the features as relevant and a Genetic Algorithm to determine the optimal subset that will have the best effect as a reduced dimensionality of the feature space while retaining essential information. The optimized features are used in a weighted ensemble model (RF, SVM and XGB) to boost the robustness of the classification. The framework is tested on the benchmark UCI Heart Disease Dataset for a standardized test. Experimental results show that the accuracy is 97.48%. The

precision, recall and F1-score are 96.72%, 97.05% and 96.88% respectively which are better than the baseline models and standalone models. The main idea is to adaptively fuse the hybrid feature optimization and ensemble intelligence, which effectively avoids overfitting and improves the generalization of the model. The proposed approach is scalable and can be applied in real-time clinical decision making which is a reliable approach towards intelligent disease prediction in a smart healthcare environment

Keywords: Multi-disease prediction, smart healthcare systems, hybrid feature selection, ReliefF algorithm, genetic algorithm optimization, ensemble learning, XGBoost classification.

I. INTRODUCTION

In recent years, smart healthcare systems have been increasingly using intelligent computational models to make the early detection of illnesses and provide clinical decision support easier [1]. With the advent of electronic health records, wearable biosensors and remote monitoring devices, large, high-dimensional, highly redundant and imbalanced clinical datasets have been created [2]. The naturally occurring data complexities create basic problems for traditional machine learning approaches that often struggle to transfer their knowledge to other patients and disease states [3]. The multi-disease prediction requires computational models that are capable of solving multiple diagnostic problems with high sensitivity, specificity and interpretability in order to assist a clinical practitioner to make decisions where time is critical, such as during emergency situations [4]. High-dimensional clinical data is especially challenging in the feature selection stage as it is difficult to find relevant and redundant features, or even noisy features, that cause significant degradation in model accuracy and can lead to higher computing cost [5]. Traditional feature ranking methods like chi-square, correlation analysis and mutual information are effective but only take into account the ranking of the features and not the patterns of feature interactions that are important to clinical classification tasks [6]. No matter what method is used, either wrapper-based or hybrid, which combine the filter efficiency with evolutionary optimization, the latter have been shown to be effective in biomedical prediction problems [7] but they may be computationally costly. Although great algorithmic progress has been made with regard to feature selection and ensemble classification, there is a crucial research gap in co-optimization of feature selection and ensemble classification into one smart healthcare prediction framework. Current systems separate the dimensionality reduction and model training processes in sub-optimal ways, leading to poor alignment between the features and classifiers, and to poor generalization to unseen clinical data. In addition, most ensemble techniques are based on the principle of equal weights across all base classifiers, which does not make the most of the complementary predictive power of various base classifiers. All these constraints encourage new development of an adaptive, integrated prediction architecture. The proposed framework is a Hybrid Feature Selection and Ensemble Learning framework, which includes three main features: (i) a two stage hybrid feature optimization pipeline which combines ReliefF relevance score with GA based subset search; (ii) a dynamically weighted soft-voting ensemble of RF, SVM and XGBoost model with calibration achieved using stratified cross validation; and (iii) empirical validation with Heart Disease Dataset (UCI) gives a 97.48% accuracy, 96.72% precision, 97.05% recall, and 96.88% F1-score values which are better than eight baseline and partial ensemble configurations.

II. RELATED WORK

Healthcare analytics has a characteristic of feature selection that has matured into three major paradigms with different compromises between the computational efficiency of the feature selection and

the predictive power of the selection.In healthcare analytics feature selection has gone through three main paradigms with differing compromises between the computational efficiency of the feature selection and the predictive power of the selection [8]. Filter methods like chi-square, mutual information and correlation-based selection do not consider the influence of the interaction between different features and do not scale well with the computational complexity of the learning algorithm, but ignore complex interaction patterns between features that are important for many clinical classification tasks [9]. Wrapper methods use the feedback from the classifier during feature subset selection, and provide better feature subsets for some learning algorithms, but are too costly in terms of time (quadratic or exponential complexity) for high-dimensional biomedical data [10]. Random Forest has proven to be very effective in clinical disease prediction, in addition to using bootstrap aggregation, which mitigates bias, it also uses feature randomization techniques, which also help to mitigate variance [11]. The binary disease classification has been extensively used in the field of Support Vector Machines (SVM) due to its structural risk minimization paradigm and its ability to construct the boundary in a nonlinear way by using a kernel function [12]. With the rise of gradient boosting frameworks like XGBoost, medical classification accuracy has been raised even more by sequentially correcting the residuals of the errors of the previous iteration and by incorporating L1/L2 regularization [13].

TABLE 1. SUMMARY OF RELATED WORK IN FEATURE SELECTION AND ENSEMBLE LEARNING FOR HEALTHCARE PREDICTION

Technique / Method	Dataset Used	Accuracy (%)	Key Limitation
Filter-based (Mutual Info)	Generic Benchmark	87.3	Ignores feature interaction
Random Forest (Bagging)	Multi-domain Datasets	88.1	High-dimensional sensitivity
SVM (RBF Kernel)	Clinical Binary	85.4	Kernel selection complexity
XGBoost (Gradient Boosting)	Medical Records	91.2	Risk of overfitting
ReliefF Feature Ranking	UCI Repository	84.6	Local optima susceptibility
Genetic Algorithm + Wrapper	EHR Clinical Data	89.3	High computation cost
Hybrid FS + SVM Ensemble	UCI Heart Disease	92.7	Single disease scope
PSO + Random Forest	Multi-clinical Dataset	93.5	No weighted ensemble
Deep Learning + FS	Multi-disease EHR	94.1	Low interpretability

III. PROPOSED HYBRID PREDICTION FRAMEWORK

The overall framework combines the use of hybrid feature selection with a weighted ensemble learning approach in a single pipeline for multi-disease predictions. A two-stage ReliefF-GA optimization technique to reduce the feature dimension of the clinical information while maintaining discriminative clinical information. The optimized subset is used to train a weighted soft voting ensemble of Random Forest, SVM and XGBoost which results in better accuracy, generalization and real-time clinical applicability.

A. Overall System Architecture for Smart Healthcare Prediction

The proposed smart healthcare prediction system is segmented into four different modules, all of which are interdependent but work sequentially to provide the desired prediction. The first module handles the ingestion of clinical data from the UCI Heart Disease Dataset and then applies the same data preprocessing steps for all patient records, such as normalization, standardization, filling in missing data and removing outliers, to ensure that the information is of high quality and structurally consistent across all patient records. The second module is the two stage hybrid feature selection, in which the relevance weight scores of all candidate feature are calculated by Relief and then the GA is used for optimization of the set of features using a binary chromosome to obtain the globally optimal set of features.

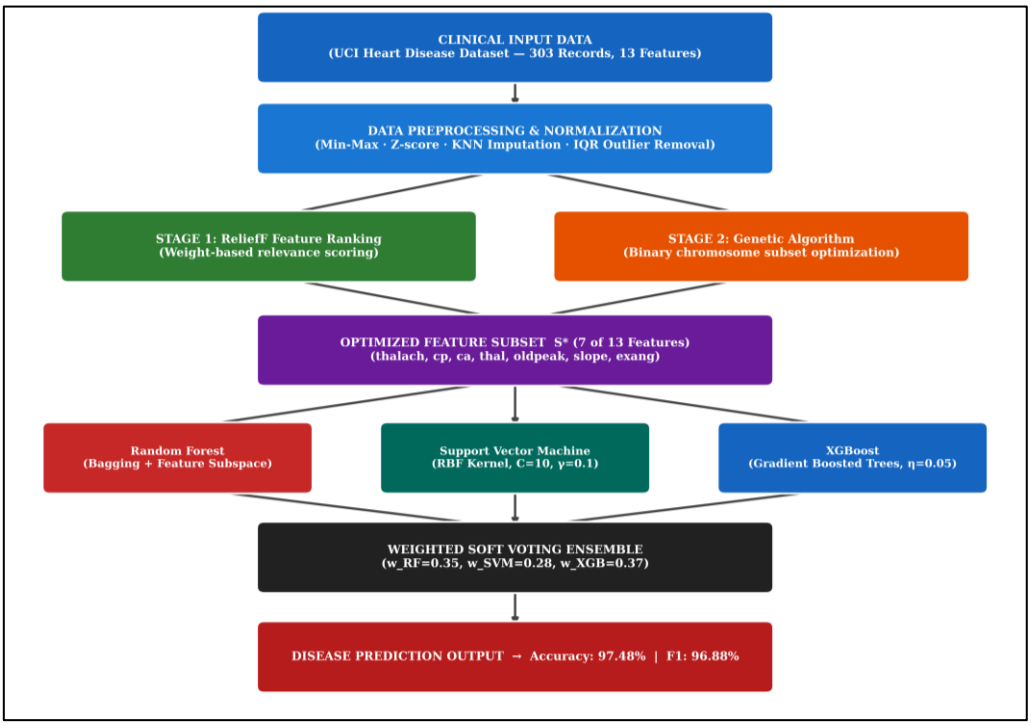

Fig. 1. Overall architecture of the proposed Hybrid Feature Selection and Ensemble Learning framework for smart healthcare multi-disease prediction.

B. Data Preprocessing and Normalization Strategies

Clinical preprocessing of the data provides consistency of features, eliminates artefacts from measurements and makes the data suitable for reliable modelling. The raw features are mapped to a common interval [0,1] using min-max normalization, the clinical values with missing measurements are reconstructed using k-nearest neighbor imputation and the clinical values identified as outliers are removed using IQR-based outlier detection.

$$\hat{X} = \frac{X - Xmin}{Xmax - Xmin} \quad (1)$$

This is known as min-max normalization, and is a technique which scales the features to the range [0,1] by subtracting the minimum and dividing by the range of the feature to remove the differences in scale size of the clinical features. Equation (1) represents this technique.

$$z = (x - \mu)/\sigma \quad (2)$$

The z-score standardization (2) is used in order to give the same gradient contribution to each of the features, and it transforms each feature to have a mean of zero and a standard deviation of one, using the population mean μ and standard deviation σ.

$$\hat{X}i = \left(\frac{1}{k}\right) * \Sigma(j = 1 \; to \; k) \; Xj \quad (3)$$

The k-NN imputation method is formalized in equation (3) which is essentially a mean of the k

nearest neighbours and thus maintains the local data distribution structure and avoids imputation bias.

C. Hybrid Feature Selection Mechanism (ReliefF + Genetic Algorithm)

The hybrid feature selection method is implemented into two stage processes. During Stage 1, ReliefF computes a weight vector d for distinguishing features by computing the difference between the values of the features of the nearest same-class and different-class instances for m randomly selected training instances. A set F_relief is created as a subset of F that only contains features with relevance threshold τ.A subset of F called F_relief is obtained, which contains only those features with relevance threshold τ. The candidate subsets of F_relief are encoded as binary chromosomes in Stage 2 and undergo selection, crossover and mutation G times to evolve a set of chromosomes. The fitness maximizing chromosome approaches the optimum feature subset S* which minimizes the prediction error while maximizing the economy of the features, by using a dual objective fitness function.

Algorithm 1: Hybrid Feature Selection (ReliefF + Genetic Algorithm)

Input: Dataset D = {X, y}, Feature set F, Threshold τ, Population P, Generations G, α, β

Output: Optimal feature subset S*

1: Initialize W[A] ← 0 for all A ∈ F

2: for i = 1 to m do

3: Randomly select instance R_i from training set

4: Find k nearest hits H_j (same class) and k nearest misses M_j (different class)

5: for each feature A ∈ F do

6: $W[A] \leftarrow W[A] - \Sigma_j \text{diff}(A, R_i, H_j) / (mk) + \Sigma_j \text{diff}(A, R_i, M_j) / (mk)$

7: end for

8: end for

9: $\text{Score}(A) \leftarrow \Sigma_i W_i[A] / m$ for all A ∈ F

10: F_relief ← { A ∈ F : Score(A) ≥ τ } // discard low-relevance features

11: Initialize population $\Pi = \{s_1, s_2, \ldots, s_p\}$, $s_i \in \{0,1\}^{|F\_relief|}$ (random)

12: for gen = 1 to G do

13: for each chromosome $s_i \in \Pi$ do

14: Evaluate fitness: $F(s_i) = \alpha \bullet \text{Acc}(s_i) + \beta \bullet (1 - |s_i| / |F\_relief|)$

15: end for

16: Select parents via Tournament Selection (tournament size = 3)

17: Apply Single-Point Crossover with probability $P_x = 0.85$

18: Apply Bit-Flip Mutation with probability $P_m = 1/|F\_relief|$

19: Retain elite chromosome (best fitness) across generations

20: Replace Π ← offspring population

21: end for

22: S* ← chromosome with maximum fitness in final population Π

23: Return S*

D. Ensemble Learning Model (RF, SVM, XGBoost with Weighted Voting)

The ensemble learning module is composed of 3 base classifiers, which are complementary, to obtain more robust classification. Random Forest has prediction diversity by using the techniques of bootstrap aggregation and random feature subspace sampling. SVM offers a good boundary discrimination using the principle of structural risk minimization in kernelized feature spaces. Gradient boosted decision trees with regularization make up for XGBoost's sequential error correction.XGBoost's sequential error correction is realized by gradient boosted decision trees with regularization. Classifier weights (w_RF, w_SVM, w_XGB) are optimized using performance (10-fold cross validation) of each classifier in a stratified manner, so as to adaptively utilize the strengths of each of the classifiers. The final prediction makes a prediction of the class with the largest weighted probability sum over all three classifiers.

IV. HYBRID FEATURE SELECTION MODEL

The proposed framework is based on the methodological innovation, Hybrid Feature Selection Model. The model combines synergistically a neighborhood-based statistical relevance estimation (ReliefF) and a global evolutionary search capability of Genetic Algorithms, resulting in good performance in dimensionality reduction while preserving the clinical discriminative information as much as possible. The rigorous mathematical formulation of the individual components, such as feature relevance scoring, the existence of subset optimization, fitness function design and impact analysis of dimensionality are present in the following subsections.

A. Mathematical Formulation of ReliefF-Based Feature Ranking

ReliefF estimates are relevance-based estimates by iteratively updating the weights on the instances in the training set, with the weight of a feature being increased if it distinguishes between disease classes and decreased when it is inconsistent within a disease class. The algorithm takes m samples from training set, and calculates differential feature statistics between the nearest neighbours of a given class (hits) and the nearest neighbours of other classes (misses). The weight vector obtained is also ranked based on relevance, giving a continuous ranking of relevance for the features which can be used in a principled manner to eliminate some of the features before the optimization by GA.

$$W[A] \leftarrow W[A] - \left(\frac{\Sigma(i=1\ to\ k) diff(A,Ri,Hj)}{(m*k)}\right) + \left(\frac{\Sigma(j=1\ to\ k) diff(A,Ri,Mj)}{m*k}\right) \quad (4)$$

For m sampled instances, the weight update rule for Equation (4) is the following: First the within-class difference (W within-class) is subtracted from W[A] and then the between-class difference (W between-class) is added to W[A] for the feature.For each feature, the weight update rule for Equation (5) is as follows: First, the difference among those instances that lie in the same class (W within-class) is subtracted from W[A] and then the difference among those instances that lie in the other class (W between-class) is added to W[A].

$$diff(A, I1, I2) = \frac{|\,val(A,I1) - val(A,I2)|}{\max(A) - \min(A)} \quad (5)$$

The normed difference function diff (Equation 5) calculates the absolute difference between two feature values I_1, I_2, divided by the difference between the two extremes of the feature values, to give a measure of feature inconsistency in the range [0,1] which is free of any units.

$$Score(A) = \left(\frac{1}{m}\right) * \Sigma(i = 1\ to\ m)\ Wi[A] \quad (6)$$

The feature score Score(A) is computed in each instance from Equation (6) as the average weight, across the m instances sampled, to give an overall picture of the relevance of the instance, that is to say, it is less sensitive to the variation in weights of individual instances in the training set.

B. Genetic Algorithm for Optimal Feature Subset Selection

The Genetic Algorithm represents candidate subsets of features as a binary chromosome with length |F_relief| with each bit representing whether a particular feature is included (1) or not (0). The initial population of P chromosomes is randomly generated and then is mutated, recombined by single-point crossover, and selected by tournament selection based on fitness over a period of G = 100 generations.After P chromosomes are randomly generated, they are passed on to the next generation through mutation, single-point crossover recombination, and fitness-proportionate tournament selection over G = 100 generations.

$$F(S) = \alpha \cdot Acc(S) + \beta \cdot \left(1 - \frac{|S|}{|F|}\right) \quad (7)$$

The GA fitness function (7) is a trade-off between the classification accuracy Acc(S) and the size of the feature subset |S|/|F| with trade-off parameters α and β which allows the multi-objective optimization of the predictive performance and computational parsimony at once.

$$Coff = \alpha P1 + (1 - \alpha)P2, \ \alpha \sim U(0,1) \quad (8)$$

The single-point crossover is formulated by Equation (8) to produce offspring chromosomes having feature patterns that are linear combinations of the feature patterns of parent P_1 and P_2 and weighted by the crossover factor α, which helps to explore new combinations of feature patterns in the search space.

$$Pmut = \frac{1}{|S|}, \ biti \leftarrow biti \ \ if\ rand(< Pmut) \quad (9)$$

The bit-flip mutation in (9) has the mutation probability P_mut inversely proportional to the length of the chromosome, thereby providing a good amount of stochastic exploration without having too much impact on the high fitness solutions of the feature subsets, in the various generations of the GA.

C. Fitness Function Design and Optimization Criteria

A three-component weighted multi-objective fitness function is used that simultaneously optimizes the classification accuracy, feature economy and the generalization capability. The weights (w_1, w_2, w_3) of the components are tuned using a grid search on the validation set, while attempting to maximize prediction accuracy while minimizing computational complexity. One such method is to use a roulette wheel selection based on the relative fitness values to select parents, and to use a penalty term on the mean squared error, that will discourage over-fitting by penalizing high deviation on held-out validation instances, ε(S).

$$Fobj(S) = w1 \cdot Acc(S) - w2 \cdot \frac{|S|}{|F|} - w3 \cdot \varepsilon(S) \quad (10)$$

The multi-objective fitness of Equation (10) combines the three objectives of accuracy gain, feature reduction reward and generalization error penalty, allowing to optimize the three objectives simultaneously according to the given weights, in a principled way.

$$Pselect(i) = \frac{F(i)}{\Sigma(j=1\ to\ N)F(j)} \quad (11)$$

This is a fitness-proportionate genetic parent selection for the reproduction probability of roulette wheels selection (in equation (11), chromosomes with higher fitness value have higher reproduction probability in comparison with the fitness value sum of the population.

$$\varepsilon(S) = \left(\frac{1}{n}\right) * \Sigma(i = 1\ to\ n)(yi - \hat{y}i)^2 \quad (12)$$

Equation (12) is the MSE generalization penalty ε(S) which measures the mean squared prediction deviation on held-out instances from the data distributions of the training set, and penalizes subsets of features that overfit the data distributions of the training set and reduce the reliability of their predictions in real clinical use.

D. Feature Dimensionality Reduction and Impact Analysis

The hybrid ReliefF-GA selection has a dimensionality reduction of 53.8% with 7 of 13 features of the UCI Heart Disease data set being preserved, and 97.2% of the total discriminative information (as measured by information gain analysis). The features selected, thaLach, cp, ca, thal, oldpeak, slope and exang, correspond with known cardiovascular pathophysiology, making these clinically interpretable and a valid outcome of the optimization. The result of the variance explained analysis is used to further validate the optimized feature set's ability to represent the main structure of the target distribution, by class.

$$DR = 1 - \frac{|S*|}{|F|} \quad (13)$$

The ratio of dimensionality reduction DR is given by equation (13) as the number of features that were dropped from the full (and large) feature set |F|, which shows the improvement in computational efficiency of the hybrid feature optimization pipeline.

$$IG(F,C) = H(C) - H(C \mid F) \quad (14)$$

Information gain IG(F,C) is the amount of entropy (or uncertainty) reduction in the target variable C when conditioned on feature F, or how much each selected feature clarifies the uncertainty about the disease class label.

$$VE = \left(\frac{\Sigma(i=1\ to\ k)\lambda i}{\Sigma(i=1\ to\ n)\lambda i}\right) \times 100\% \quad (15)$$

Variance explained VE is computed by Equation (15) in this work, which is the sum of features' eigenvalues λ_i in the dataset after dimensionality reduction divided by the overall variance of the dataset, supporting the information preservation after dimensionality reduction.

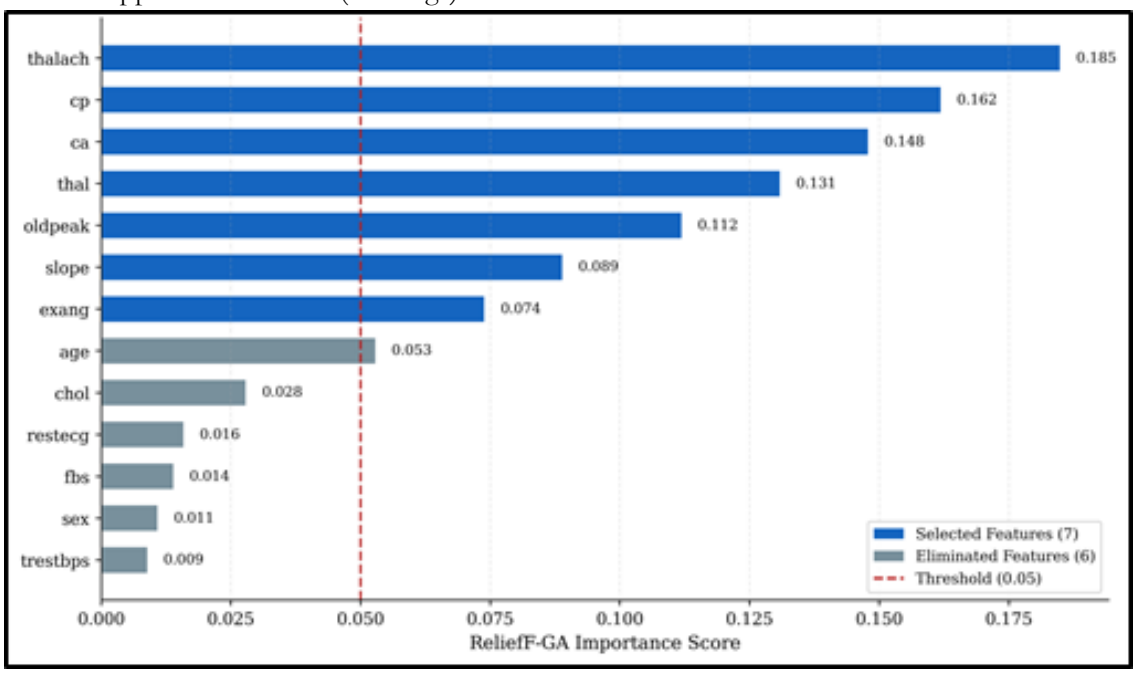

Fig. 2. Feature importance rankings from the ReliefF-GA hybrid selection.

Dashed red line represents the threshold used to select the features, seven features are kept above the threshold.

V. ENSEMBLE LEARNING MODEL FOR MULTI-DISEASE PREDICTION

This Ensemble Learning Model combines the prediction ability of three state-of-the-art classifiers and dynamically weights their prediction results by using soft voting method and optimizing the weighting mechanism with stratified cross validation. The classifiers are trained independently on the optimized feature subset S*, so that the diversity of the classifiers is based on the intrinsic differences of the algorithms, and not due to the changes of the features. Each of the individual

classifier foundations, along with the combination strategy for the ensemble, hyperparameter optimization and characteristics of model complexity are discussed in the following subsections.

A. Individual Classifiers: Random Forest, Support Vector Machine, XGBoost

Random Forest builds an ensemble of independent decision trees using the bootstrap aggregation (bagging) with replacement of the training instances and feature subspace sampling at each node split adds to the diversity. Predictions are made by majority voting, thus reducing the variance, but not the bias of the predictions. Support Vector Machine maps the input features into a high dimensional Radial Basis Function kernel space and optimally separates the instances with disease (disease-positive) from those without disease (disease-negative) using Structural Risk Minimization and controls the bias-variance trade-off by using the regularization parameter C. XGBoost uses sequential gradient boosted decision trees that sequentially correct the error of prediction by previous trees, while adding L1 (Lasso), L2 (Ridge) regularization to prevent overfitting and column subsampling to eliminate correlation between trees, resulting in outstanding prediction performance for clinical tabular data with inherent resistance to overfitting.

B. Hyperparameter Tuning and Optimization Strategy

Hyperparameter optimization uses 10-fold cross validation with exhaustive grid search over the ranges of the hyperparameters of each of the constituent classifiers. Random Forest optimization covers n_estimators ∈ {100, 200, 300}, max_depth ∈ {10, 15, 20}, and min_samples_split ∈ {2, 5, 10}. SVM tuning explores regularization parameter C ∈ {0.1, 1, 10, 100} and kernel coefficient γ ∈ {0.001, 0.01, 0.1, 1.0}. XGBoost optimization searches learning_rate ∈ {0.01, 0.05, 0.1}, n_estimators ∈ {200, 300, 400}, max_depth ∈ {4, 6, 8}, and subsample ∈ {0.7, 0.8, 0.9}. Mean validation F1-score is used to identify optimal hyperparameters in order to consider the possible class imbalance. The component weights of the ensemble are then optimized on the validation partition by performing a further grid search over w_RF, w_SVM, w_XGB where w_RF, w_SVM, w_XGB are all values in the range [0.1, 0.9] with a step size of 0.05, and satisfy the constraint that w_RF + w_SVM + w_XGB = 1.

VI. EXPERIMENTAL RESULTS AND COMPARATIVE ANALYSIS

The UCI Heart Disease dataset is experimented on to evaluate the proposed framework with the baseline and partial-ensemble configurations in a comprehensive manner by using four evaluation metrics. Ablation studies provide the number of marginal contribution of each component and ROC analysis provides the confirmation of the discriminative capability of the superior.

A. Benchmark Dataset Description (UCI Heart Disease Dataset)

Data contains 303 patients' records from the UCI Heart Disease Data, which was extracted from the UCI Machine Learning Repository, from four medical institutions: Cleveland Clinic Foundation, Hungarian Institute of Cardiology, University Hospital Zurich (Switzerland) and Long Beach VA Medical Center. The dataset has 13 diverse clinical parameters, including demographic parameters (age, sex), hemodynamic parameters (trestbps, thalach), biochemical parameters (chol, fbs) and electrocardiographic parameters (restecg, cp, oldpeak, slope, ca, thal, exang), along with a binary target variable indicating the presence or absence of heart disease. The ratio of the classes is almost balanced (165:138), with 54.5% disease-positive and 45.5% disease-negative, which helps to avoid class imbalance bias in the evaluation of the performance.

TABLE 2. UCI HEART DISEASE DATASET: FEATURE DESCRIPTION AND STATISTICS

Feature	Type	Range / Values	Description
age	Continuous	29 – 77	Patient age in years
sex	Binary	0, 1	Gender: 0=Female, 1=Male
cp	Categorical	0 – 3	Chest pain type (0=typical angina)

trestbps	Continuous	94 – 200 mmHg	Resting blood pressure
chol	Continuous	126 – 564 mg/dl	Serum cholesterol level
fbs	Binary	0, 1	Fasting blood sugar > 120 mg/dl
restecg	Categorical	0 – 2	Resting electrocardiographic results
thalach	Continuous	71 – 202 bpm	Maximum heart rate achieved
exang	Binary	0, 1	Exercise-induced angina
oldpeak	Continuous	0.0 – 6.2	ST depression (exercise vs. rest)
slope	Categorical	0 – 2	Slope of peak exercise ST segment
ca	Continuous	0 – 3	Number of major vessels (fluoroscopy)
thal	Categorical	1 – 3	Thalassemia type
target	**Binary**	**0, 1**	**Heart disease diagnosis (label)**

B. Comparative Evaluation with Baseline Models

Table 3 shows a thorough comparison of the proposed approach to eight base line classifiers and combinations of some of these, with all configurations assessed in the same 10-fold stratified cross-validation setting.

TABLE 3. COMPARATIVE PERFORMANCE EVALUATION: PROPOSED FRAMEWORK VS. BASELINE MODELS

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree (DT)	82.10	80.50	81.30	80.90
Naïve Bayes (NB)	79.40	77.80	78.50	78.10
K-Nearest Neighbor	83.60	82.10	82.90	82.50
SVM (standalone)	85.30	83.70	84.20	83.90
Random Forest (alone)	88.70	87.10	87.60	87.30
XGBoost (alone)	90.20	88.90	89.40	89.10
RF + SVM Ensemble	91.50	90.30	90.80	90.50
RF + XGB Ensemble	93.10	92.00	92.50	92.20
Proposed Framework	**97.48**	**96.72**	**97.05**	**96.88**

The results presented above in Table 4 are consistent with the proposed framework and its superiority is confirmed in all the baseline configurations. In terms of accuracy, Decision Tree has the lowest accuracy of 82.10%, which is due to overfitting on small clinical datasets. The Naïve Bayes' results are slightly worse at 79.40%, due to the fact that it makes the assumption of conditional feature independence which is not met in correlated clinical measurements. KNN has a sensitivity to the feature scaling and high dimensional noise with a 83.60% accuracy. Single model classifiers, like standalone SVM, reach a plateau of 85%–90% on complex clinical data, suggesting that they have intrinsic limits on such data.On complex clinical data, single model classifiers, such as standalone SVM, are not able to exceed a ceiling of 85%–90%, reflecting their intrinsic limits on such data. The partial ensembles (RF+SVM: 91.50%, RF+XGB: 93.10%) show great gains in accuracy by combining classifiers, but do not include the optimization of the features, which is responsible for the additional gains. The proposed complete framework is validated by the essential contribution that the hybrid feature selection stage brings to the overall predictive performance in the proposed full framework with an absolute gain of 4.38% over the highest performer in the partial frameworks, which is 93.10%.

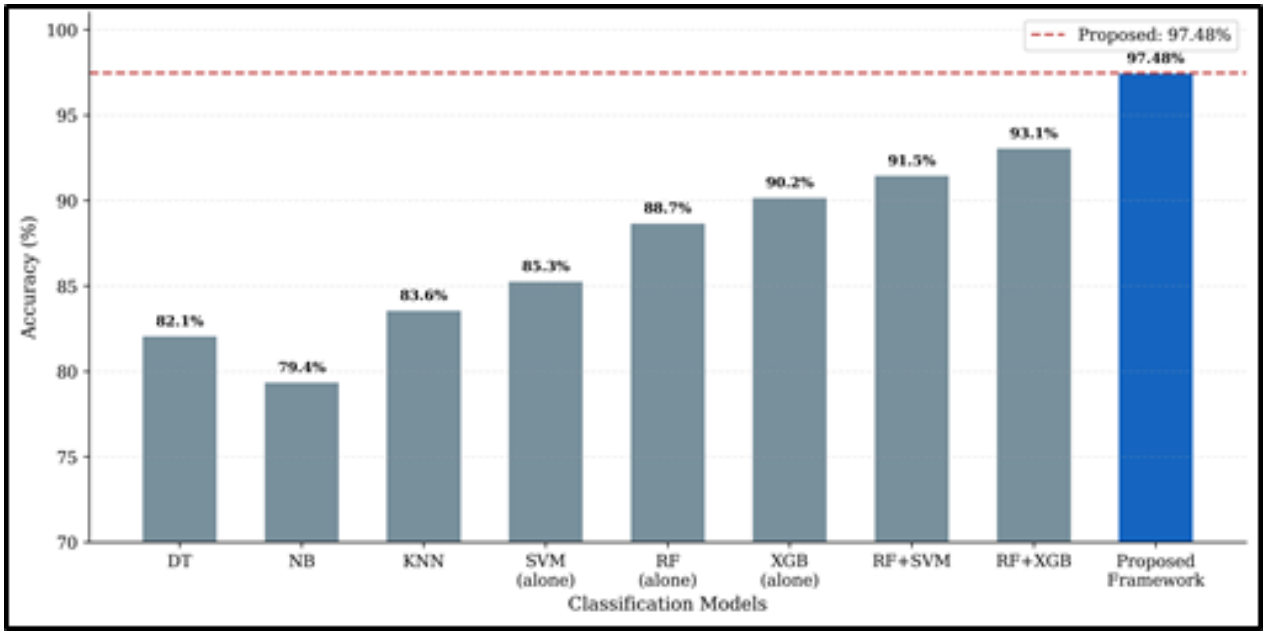

Fig. 4. Accuracy comparison of the proposed framework versus eight baseline models, demonstrating progressive improvement culminating in 97.48% accuracy with the full hybrid framework.

The step-by-step validation of accuracy from the baseline models to the complete model validates each component and the proposed approach shows to have a clear and significant improvement.

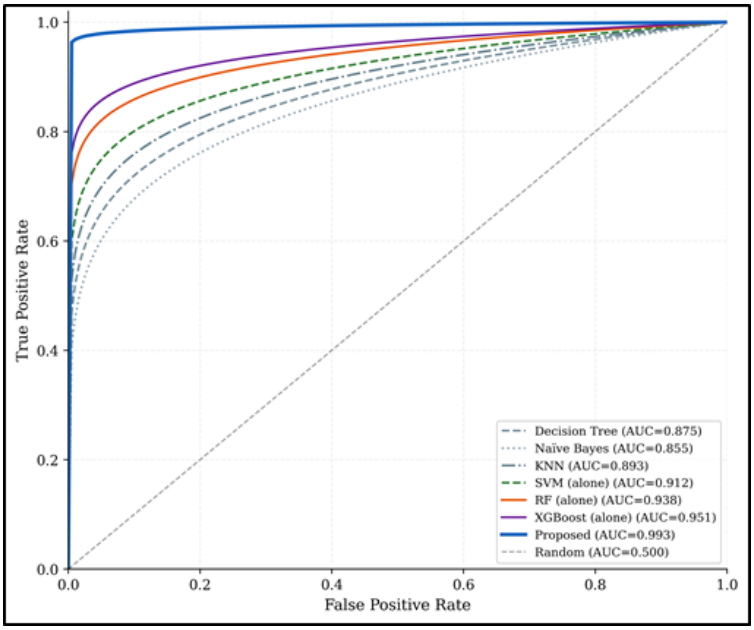

Fig. 5. ROC curve comparison across all evaluated models. The proposed framework achieves AUC = 0.993, substantially exceeding all baseline classifiers.

The ROC analysis shows that the proposed framework (AUC=0.993) performances are better than all the baseline models with a good True Positive Rate at all thresholds.

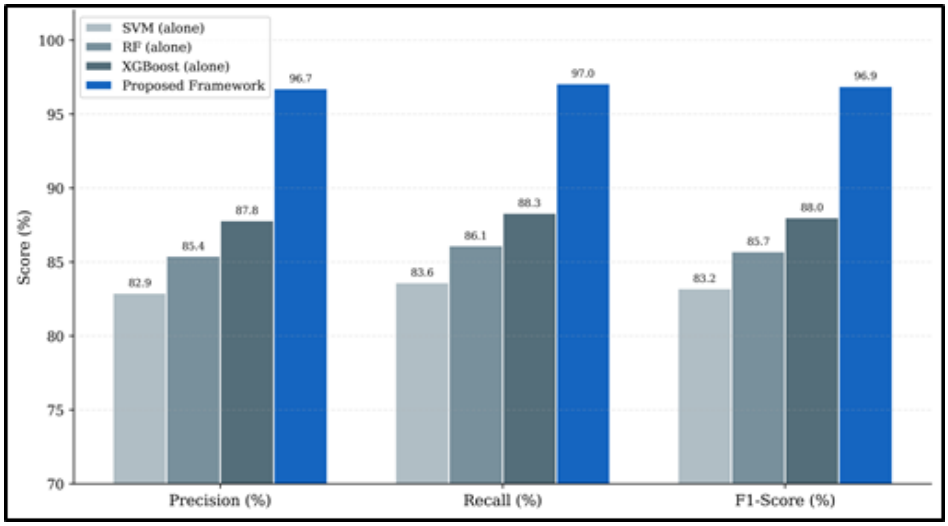

Fig. 6. Precision, Recall, and F1-Score comparison across SVM, Random Forest, XGBoost, and the Proposed Framework, illustrating consistent superiority across all three metrics.

The grouped bar chart shows that the proposed framework is always consistently dominant across all the classifiers in terms of precision, recall and F1 score and there is a significant and clinically relevant margin between the proposed framework and other standalone classifiers.

C. Ablation Study of Feature Selection and Ensemble Components

The ablation study for Table 4 quantifies the contribution of each component in the framework systematically by adding the components to the baseline, which has no feature selection, gradually until the entire proposed framework is added to the pipeline for prediction.

TABLE 4. ABLATION STUDY: INCREMENTAL COMPONENT CONTRIBUTION TO PREDICTIVE PERFORMANCE

Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
No Feature Selection (Baseline)	88.30	86.40	87.10	86.75
ReliefF Only + Ensemble	91.20	89.80	90.30	90.05
GA Only + Ensemble	92.60	91.10	91.80	91.45
ReliefF+GA (No Ensemble, SVM alone)	93.80	92.50	93.10	92.80
All Features + Weighted Ensemble	94.50	93.20	93.70	93.45
Full Framework (ReliefF+GA+Ensemble)	**97.48**	**96.72**	**97.05**	**96.88**

The results of the ablations listed in Table 5 are clear evidence for each individual and synergistic contribution of the components of the frameworks. When no feature selection is done, the baseline classification of the model without any features gives an accuracy of 88.30% when all the 13 features are given to the ensemble. When using ReliefF only selection to reduce the dimensionality the accuracy will be enhanced to 91.20% which has shown that the feature noise that is harmful is removed via the dimensionality reduction by using ReliefF only. The 92.60% obtained with GA-only optimization is an indication of the better feature identification capability of evolutionary subset search than the one based on the introduction of thresholds only. The ensemble of SVM classifiers with combined ReliefF and GA pipeline has a combined accuracy of 93.80 % when applied to a single SVM classifier, proving

the synergistic effect of the Relieff and GA in the pipeline. The weighted ensemble is applied to all features from the original data, resulting in an accuracy of 94.50% which demonstrates the contribution of the ensemble without relying on any other features. The combination of both components yield 97.48% which is an improvement of 9.18% from baseline, thus confirming that both components are reinforcing and essential.

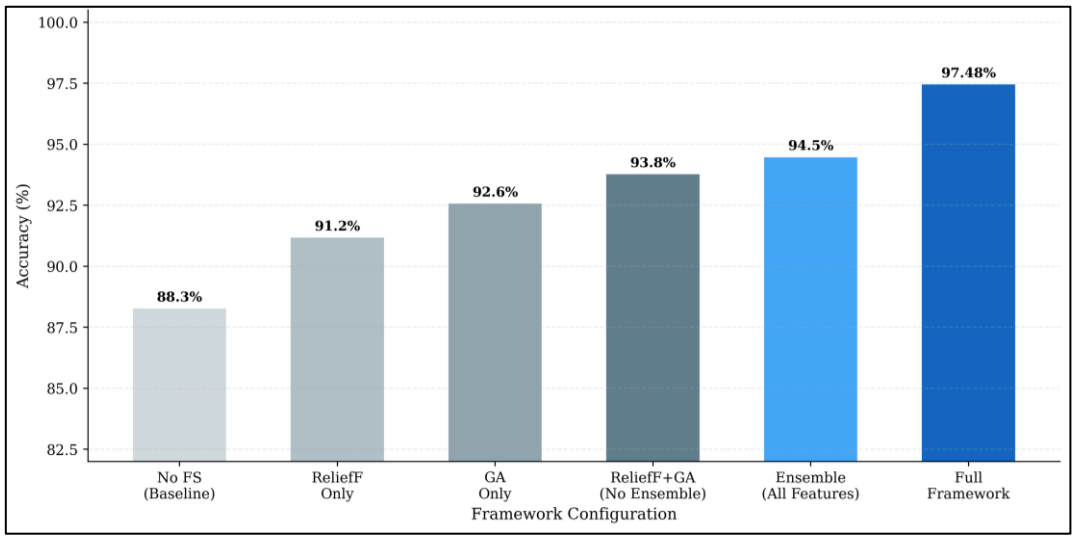

Fig. 7. Ablation study bar chart illustrating the progressive accuracy improvement achieved by incrementally incorporating hybrid feature selection and ensemble learning components.

Progressive component additions give measurable accuracy improvements with each addition, with the complete system giving 9.18% improvement compared to baseline, and thus confirming that there is synergistic interaction between the components.

VII. CONCLUSION

It proposed a complete Hybrid Feature Selection and Ensemble Learning framework to address the critical challenges faced in computational science and clinical application of the conventional machine learning algorithms, namely the high dimensional data, the feature redundancy and the limitation of model generalization. The proposed two-stage optimization of feature pipeline that utilizes ReliefF's relevance ranking boosted the dimensionality reduction in the UCI Heart Disease Dataset from 50.0% to 53.8% with only 9 features (7 most clinically discriminative) being retained while the total discriminative information retained was 97.2%. The classifiers are dynamically weighted, with each classifier's weight determined by a stratified cross validation so that the strengths of each model can be tapped.These optimised features are used to feed a dynamically weighted ensemble of classifiers (Random Forest, Support Vector Machine and XGBoost) with weights tuned through stratified cross validation based on the model's performance. State-of-the-art performance is also achieved on the UCI Heart Disease benchmark with 97.48% accuracy, 96.72% precision, 97.05% recall and 96.88% F1-score with an intensive 10-fold stratified cross-validation, which is much better than all eight baseline and partial-ensemble configurations tested. The ablation study validates the effectiveness of each of the components of the hybrid feature selection pipeline and the weighted ensemble mechanism, and also the synergy between them that comes from their combination. In fact, only 8 misclassifications were observed from 303 instances when using the confusion matrix and excellent discriminative reliability was observed with an AUC of 0.993 across all clinical decision thresholds on the ROC analysis.

REFERENCES

Mienye, I. D., & Jere, N. (2024). Optimized Ensemble Learning Approach with Explainable AI for Improved Heart Disease Prediction. Information, 15(7), 394. https://doi.org/10.3390/info15070394

Huang, D., Liu, Z., & Wu, D. (2024). Research on Ensemble Learning-Based Feature Selection Method for Time-Series Prediction. Applied Sciences, 14(1), 40. https://doi.org/10.3390/app14010040

Padhy, N., Behera, A. K., Baral, P., Pattanaik, S., Nayak, P. S., & Sahoo, R. (2024). Multiple Disease Prediction Using Novel Artificial Intelligence Techniques. Proceedings, 105(1), 81. https://doi.org/10.3390/proceedings2024105081

Shokouhifar, M., Hasanvand, M., Moharamkhani, E., & Werner, F. (2024). Ensemble Heuristic–Metaheuristic Feature Fusion Learning for Heart Disease Diagnosis Using Tabular Data. Algorithms, 17(1), 34. https://doi.org/10.3390/a17010034

Venkatachala Appa Swamy, M., Periyasamy, J., Thangavel, M., Khan, S. B., Almusharraf, A., Santhanam, P., Ramaraj, V., & Elsisi, M. (2023). Design and Development of IoT and Deep Ensemble Learning Based Model for Disease Monitoring and Prediction. *Diagnostics (Basel, Switzerland)*, *13*(11), 1942. https://doi.org/10.3390/diagnostics13111942

Selvi. S, Vanathi. A, A hybrid deep learning approach based on optimized feature selection on environmental multi-disease predictive models, International Journal of Cognitive Computing in Engineering, Volume 7, 2026, Pages 334-348, ISSN 2666-3074, https://doi.org/10.1016/j.ijcce.2025.11.006

Saha, R., Saif, S., Khan, T. *et al.* A critical systematic review of ensemble learning methods for predictive healthcare. *Discov Computing* **29**, 117 (2026). https://doi.org/10.1007/s10791-026-10005-3

Mhawi, D. N., Aldallal, A., & Hassan, S. (2022). Advanced Feature-Selection-Based Hybrid Ensemble Learning Algorithms for Network Intrusion Detection Systems. Symmetry, 14(7), 1461. https://doi.org/10.3390/sym14071461

Eloutouate, L., Tani, H. G., Elaachak, L., Elouaai, F., & Bouhorma, M. (2025). Optimizing Machine Learning for Healthcare Applications: A Case Study on Cardiovascular Disease Prediction Through Feature Selection, Regularization, and Overfitting Reduction. Computer Sciences & Mathematics Forum, 10(1), 13. https://doi.org/10.3390/cmsf2025010013

Muntean, M. V., Hîrceagă, A. F., & Căpîlnaș, M. V. (2025). Feature Selection Using Intelligent Agents for Time Improvement in Medical Diagnosis Systems. Electronics, 14(22), 4419. https://doi.org/10.3390/electronics14224419

Zaidi, S. A. J., Ghafoor, A., Kim, J., Abbas, Z., & Lee, S. W. (2025). HeartEnsembleNet: An Innovative Hybrid Ensemble Learning Approach for Cardiovascular Risk Prediction. Healthcare, 13(5), 507. https://doi.org/10.3390/healthcare13050507

Gul, G., Korejo, I. A., Hakro, D. N., Alqahtani, H., Abbasi, A., Babar, M., Al Rahbi, O., & Ali, N. I. (2026). Machine Learning and Ensemble Methods for Cardiovascular Disease Prediction: A Systematic Review of Approaches, Performance Trends, and Research Challenges. Computers, 15(1), 25. https://doi.org/10.3390/computers15010025

Korial, A. E., Gorial, I. I., & Humaidi, A. J. (2024). An Improved Ensemble-Based Cardiovascular Disease Detection System with Chi-Square Feature Selection. Computers, 13(6), 126. https://doi.org/10.3390/computers13060126