Advances in Consumer Research

https://acr-journal.com/

Volume-2 | Issue-6 | Dec 2025

Original Researcher Article

Hybrid Ensemble Architecture with Feature Selection for Predicting Risk of Cardiovascular Disease

Sanjeev Bhardwaj[1] and Deepankar Bharadwaj[2]

[1,2]Department of Computer Science and Engineering, IFTM University, Lodhipur Rajput, Moradabad, 244102, UP, India.

Received: 11/11/2025
Revised: 30/11/2025
Accepted: 15/12/2025
Published: 27/12/2025

ABSTRACT

Cardiovascular disease (CVD) remains a significant worldwide health challenge, thereby necessitating precise and dependable analytical models for early risk estimation. Existing machine learning methodologies frequently exhibit performance decay restricting from the presence of redundant or irrelevant clinical features. This research familiarizes an enhanced predictive framework, incorporating Chi-Square feature selection combined with an ensemble-based classification model, with the objective of improving diagnostic accuracy and computational efficiency. The Chi-Square technique is utilized to determine statistically important attributes that contribute to identify CVD risk, thus facilitating dimensionality decrease and strengthening model interpretability. The preferred features are subsequently employed to train an improved ensemble model, which integrates Random Forest, Gradient Boosting Machine (GBM), and Extra Trees Classifier, and combines them through a soft-voting methodology. Model estimation is conducted using accuracy, precision, recall (Sensitivity), F1-score and AUC-ROC metrics to simplify a thorough assessment of performance. Initial findings recommend that the proposed Chi-Square boosted ensemble architecture surpasses predictable single classifiers and ensemble models missing feature selection, exhibiting greater predictive solidity and moderated overfitting. This study ultimately suggests that the combination of statistical feature selection with ensemble learning grants an added dependable and scalable approach for CVD risk prediction, by this means contributing significant improvement to computational healthcare and protective cardiology.

Keywords: Cardiovascular Disease, Ensemble Learning, Chi-Square Feature Selection, Machine Learning, Soft-Voting Classifier, Predictive Modelling, Computational Healthcare, Preventive Cardiology.

1. INTRODUCTION:

Cardiovascular diseases (CVDs) continue to be the main cause of death globally, establishing an important international health challenge. These involve coronary artery disease, stroke, heart failure, and several other conditions that impression the heart and vascular system. Recent global health data specifies that CVDs were accountable for roughly 19.8 million death rates in 2022, accounting for just about 32% of all global deaths [1]. Additionally, the occurrence of CVD-related deaths has dependably increased, from 13.1 million in 1990 to completed 19 million in 2023, thereby representative a sustained rising course [2].

This burden is suspiciously shouldered by low and middle-income countries, which denote over three quarters of all worldwide CVD related mortalities [1]. These annotations underscore the considerable socio-economic consequences of CVDs, encompassing heightened healthcare expenses, reduced productivity, and financial compressions on both peoples and national healthcare organizations. So, the early and specific assessment of CVD risk is indispensable for facilitating rapid intervention and preventative actions.

Traditional risk prediction methods, demonstrated by the Framingham Risk Score (FRS), have been essential in clinical situations. Even though the FRS and equivalent statistical models are widely employed, their support on a restricted arrangement of well-known risk factors such as precisely age, cholesterol levels, blood pressure, smoking, and diabetes. These models also presume linear relationships between the variable quantity. Accordingly, their applicability across wide-ranging populations is controlled, potentially leading to an under-estimation of risk within the groups characterized by different baseline attributes or eco-friendly inspirations [3]. Additionally, these models commonly supervise the complicated, multifactorial, and non-linear exchanges that underlie cardiovascular risk.

Machine learning based prediction models have raised as promising another possibility, demonstrating the measurements to analyse general datasets and identify complex outlines that better capabilities of predictable methods. Recent research highlights their possible to

enhance problem-solving accuracy and risk stratification [4].

On the other hand, machine learning methodologies encounter important complications. The high dimensionality inherent in clinical datasets often results in the inclusion of redundant and irrelevant features, potentially diminishing model effectiveness. Additionally, the existence of noisy, incomplete, and mixed data, frequently observed in trustworthy health records, presents additional problems during model training. Besides, numerous high performance machine learning models operate like as black boxes that providing controlled interpretability. a crucial impairment to clinical application, given the obligation of transparency and explainability [4]. These limits highlight the authoritative for enhanced outlines that take part predictive accuracy, robustness, and interpretability.

Although machine learning (ML) is vital for analysing many types of diseases, a most important problem with ML algorithms is the formation of large datasets that often include superfluous and recurring features [5]. Moreover, in numerous cases, only a few features are really important for the use in task. As a result, the performance and accuracy of the classification are negatively exaggerated when the remaining features are considered insignificant and redundant.

Therefore, it's essential to pick a small, applicable set of important features to improve classification performance and decrease the expletive of dimensionality. Feature selection techniques are used to measure how significant different features are. The primary goal is to condense the number of inputs to only those that are most appropriate to the model.

Furthermore, feature selection technique not only diminishes the input dimensionality but also significantly accelerates processing. Despite the application of several feature selection methods within health datasets for decision support systems, opportunities for improvement keep it up [6]. Previous research into heart disease prediction have largely focused on either algorithmic optimization through diverse machine learning approaches or the fine-tuning of algorithms via the application of various feature selection techniques.

India holds the second place worldwide in terms of Human Development Index (HDI) ratings. Heart disease (HD) is recognized as the main cause of mortality globally [7]. Strokes can visible suddenly, inclined by factors such as excessive alcohol consumption, obesity, poor sleep behaviors, absence of physical activities, and numerous medical surroundings such as hypertension, diabetes and cholesterol along with stress and tobacco use [8]. Several data analytics knowledges have been employed to assistance healthcare specialists in the early finding of HD, employing a restricted set of prepared parameters that recognize the underlying causes of the condition. Feature selection is a serious phase in technique improvement as it improves data by minimizing the number of participation variables.

Automating the diagnostic process offers a promising avenue for reducing healthcare costs, saving valuable time for both patients and clinicians, and improving the efficiency of clinical workflows. In this context, the present study goals to develop a precise and consistent diagnostic framework for Huntington's disease. To improve the precision and reliability of HD detection, advanced data study techniques such as feature selection and principal component analysis (PCA) are employed [9]. In addition, the integration of ensemble learning approaches is expected to further strengthen the model's diagnostic performance. Rather than replacing clinical judgment, this framework is intended to support healthcare professionals by providing an effective decision-assistance tool for diagnosis and treatment planning.

1.1 Novelty and Contributions

This research presents a complete cardiovascular disease (CVD) risk prediction framework, which integrates Chi-Square feature selection together with optimized ensemble learning algorithms specifically Random Forest, Gradient Boosting and XGBoost. This advanced integration effectively moderates dimensionality, removes irrelevant features and improves projecting precision. Additionally, a multi-stage optimization method, employing Grid Search, Random Search, and Bayesian Optimization, contributes to enhanced model stability and provide overall performance. Through the assessment of various ensembles on a consistent Chi-Square filtered feature set, the study provides a balanced comparative analysis, thereby showing superior accuracy, efficiency, and interpretability. Accordingly, the developed model makes available a clinically appropriate, robust, and computationally well-organized method for the early prediction of CVD risk.

The rest of this study will be prearranged as follows: The "Literature Survey" section reviews the relevant literature. The "Methodology" section defines which types of methods used in research. The "Results" section provides the research findings. After that, the study's outcomes are discussed in the "Discussion" section. the "Conclusion" section summarizes the whole work done. Finally, the "Limitation and future scope" section talks about the limitations in this work and suggests areas for future research.

2. Literature Survey:

The Cleveland UCI dataset contains numerous studies focused on predicting heart disease. These studies can be generally categorized into two main zones: one that compares algorithms based on traditional or deep learning methods, and another that estimates algorithms that use feature selection.

Recently, various machine learning methods have been developed for identifying CVD. In addition, artificial intelligence (AI) is also used in several medical fields, such as Radiology, Dermatology, Hematology and Ophthalmology. Heart disease (HD) is the main source of mortality in India, with a study signifying a 29.4% occurrence among adults aged 45 and above. The primary risk factors involve like as age, gender, geographical location, cholesterol levels, diabetes, inactive lifestyles, and family tendency, thereby highlighting the requirement for focused health creativities and proactive screening strategies. Additionally, while effective management has demonstrably better-quality the forecast for individuals with hemophilia, their advancing age correlates with heightened vulnerability to HD, hence requiring a more comprehensive knowledge and management method for this demographic [10].

Jian Ping [11] introduced a new feature selection algorithm considered for diagnosing HD, applying a series of classification algorithms. Preprocessing methods, including Relief, LASSO, mRMR and LLBFS shaped high precision and the model's performance was improved, leading to enhanced computational effectiveness. The achievement of the projected Fast Provisional Mutual Statistics feature selection algorithm was then associated to that of additional algorithms. Once this applied on the Cleveland dataset, the SVM, in aggregation with the planned method, accomplished an accuracy is 92.37%. However, the potential exists for a comprehensive decision-making outline for HD, encompassing both treatment and control measures.

Many researches are presently underway, meant at identifying effective medical diagnostic methodologies appropriate to a range of ailments. This specific study employs classification techniques to simplify efficient diagnostic predictions, utilizing a condensed set of attributes that are most influential in the context of cardiovascular illness. Chen et al. [12] expressed a breast cancer finding model by using a support vector machine (SVM) and an irregular set-based feature selection method. Wang et al. [13] executed linear kernel SVM classifiers for heart disease HD finding, accomplishing an accuracy rate of 83.37%. Furthermore, a hybrid neural network approach was presented in ref. [14], with a described accuracy of 86.8%.

A stacked ensemble classifier, which fit in several machine learning techniques, is presented and representing an accuracy is 92.34%, thereby exceptional the performance reported in previous studies. Raza [15] studied heart disease recognition through ensemble learning model with majority voting strategies and utilizing clinical type reports. The findings discovered that the voting model reached an accuracy of 89% when compared to individual classification models, suggesting its likely utility in U-healthcare monitoring structures to improve finding and decision-making processes in cardiovascular disease care. Mienye et al. [16] fixated on the development and application of an enhanced ensemble learning technique for predicting risk of heart disease. This methodology contains breakdown the dataset into reduced subsets through a mean-based breaking technique followed by modelling, this subdivision using classification and regression tree techniques.

The model established higher performance compared to previous research, getting classification accuracies of 93% and 91% when evaluated on the HDDC and FHSD respectively. Ambrews et al. [17] employed stacking and voting methodologies for HD forecast across numerous datasets. This investigation pursues to offer a short assessment of ensemble learning's success in refining predictive accuracy for cardiac disease finding. Voting is between all classifiers, produced significant outcomes on the UCI heart disease dataset (UHDD), with an accuracy of 91.96%, a F1-score of 91.69%, a recall of 91.72%, a precision of 92.40% and a specificity of 90.77%.

In summary, while existing approaches demonstrate promising performance, they also reveal that the inherent complexity of real-world clinical data cannot be adequately addressed by a single method. This underscores the need for models that achieve both high accuracy and strong generalizability through the integration of effective optimization strategies, robust feature selection, and ensemble learning techniques. Motivated by these observations, the researchers propose a hybrid feature selection framework that combines the strengths of individual feature-based methods with ensemble-based approaches to more effectively capture the underlying patterns in clinical data.

3. Materials and methods:

A complete summary of the methodology that was applied in this research is comprised in this section. It includes information about the dataset that was utilized, the preprocessing procedures, the feature selection approaches, the classifier selection technique, and the ensemble learning technique.

3.1 Proposed Methodology:

3.1.1 Dataset

A compilation of five highly-regarded datasets—the Long Beach VA, Switzerland, Hungarian, Statlog, and Cleveland datasets—this extensive collection has 1,190 records with 11 attributes [18]. Below in Figure 2, you can find a brief description of this dataset that was published by Alizadehsani et al. The data type and description of each of the 11 features are demonstrated in the Figure 1.

S. No.	Features	Description	Data Type
1	Age	The age of the patient (29–77)	Integer
2	Sex	Male=1, Female=0	Binary
3	Chest pain type	Various type of pain in chest	Nominal
4	Resting bp s	blood pressure at rest	Integer
5	Cholesterol	Cholesterol levels range from 126 to 564 mg/dl	Integer
6	Fasting blood sugar	Blood sugar level during fasting > 120 mg/dl, yes = 1, false = 0	Nominal
7	Resting ecg	ECG Resting & Electrocardiographic resting result (0–1)	Numeric
8	Max heart rate	Heart rate: 60 at minimum, 202 at maximum	Integer
9	Exercise angina	Workout comprised angina (correct = 1, nothing = 0)	Nominal
10	Oldpeak	Exercise-induced ST depression in comparison to rest (0–2)	Integer
11	ST slope	The highest point of the workout ST segment's slope (0–1)	Nominal
14	Target	Cardiac illness = 1, No cardiac illness = 0	Binary

Figure 1: Attribute description of heart_statlog_cleveland_hungary_final dataset

	age	sex	chest pain type	resting bp s	cholesterol	fasting blood sugar	resting ecg	max heart rate	exercise angina	oldpeak	ST slope	target
0	40	1	2	140	289	0	0	172	0	0.0	1	0
1	49	0	3	160	180	0	0	156	0	1.0	2	1
2	37	1	2	130	283	0	1	98	0	0.0	1	0
3	48	0	4	138	214	0	0	108	1	1.5	2	1
4	54	1	3	150	195	0	0	122	0	0.0	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...
1185	45	1	1	110	264	0	0	132	0	1.2	2	1
1186	68	1	4	144	193	1	0	141	0	3.4	2	1
1187	57	1	4	130	131	0	0	115	1	1.2	2	1
1188	57	0	2	130	236	0	2	174	0	0.0	2	1
1189	38	1	3	138	175	0	0	173	0	0.0	1	0

1190 rows × 12 columns

Figure 2: Sample data with all relative features.

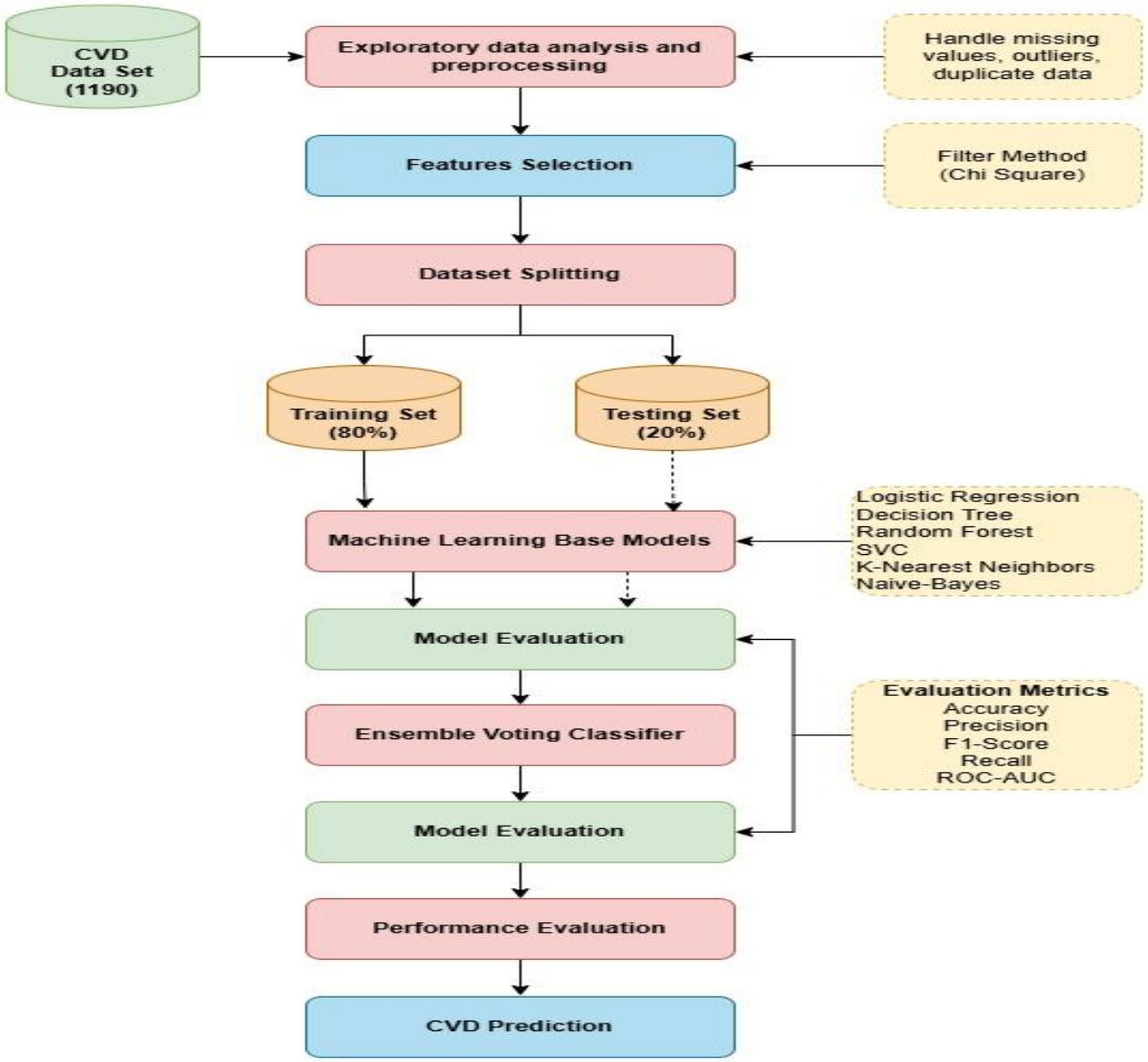

Figure 3: An outline of the overall workflow of the proposed system.

3.2 Data Preprocessing

Prior to the implementation of the projected algorithm, data preprocessing is important to assurance the providing of high-quality inputs for the model. This preprocessing phase helps to eliminate noise, correct inconsistencies and normalize the data, in that way enhancing the prediction model's performance and consistency. The primary preprocessing procedures employed in this investigation involve outlier detection and removal. applying the Inter-Quartile Range (IQR) method, succeeded by data normalization to ensure all features are scaled comparably. These measures are particularized upon in the subsequent sections. Figure 4 shows the distribution of target variables in dataset.

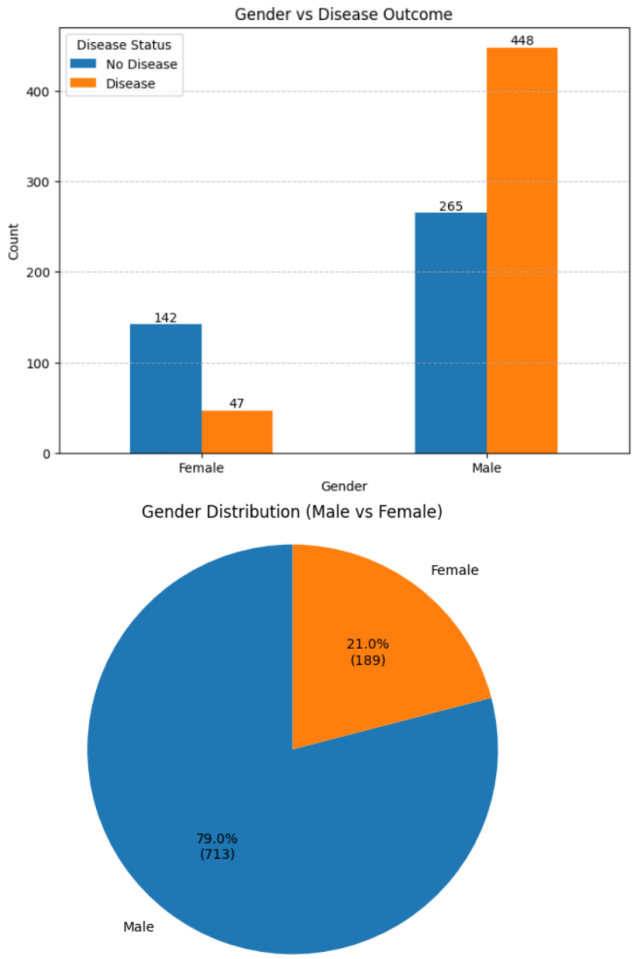

Figure 4: Distribution of target variables in dataset (Gender distribution with disease).

3.2.1 Handling Missing Values

If missing values are not correctly addressed then the performance of the model can suffer. Mutual imputation strategies for numerical features include mean imputation, which is suitable for data that is generally distributed and median imputation. which is healthy to twisted distributions and outliers. Mode imputation is classically applied for categorical variables. In order to sustain data consistency, missing values are exchanged with statistical measures [19].

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (1)$$

Each missing value ($\boldsymbol{x_{miss}}$)is substituted by ($\bar{\boldsymbol{x}}$)

Median Imputation:

$$\boldsymbol{x_{miss}} = \text{median}(\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}) \quad (2)$$

More robust when the data holds outliers.

Mode Imputation (Categorical features):

$$x_{miss} = \arg\max_{c} \; ; \text{count}(c) \quad (3)$$

3.2.2 Feature Scaling / Normalization

Feature scaling is vital for ensuring that numerical variables have an equivalent effect on model training, a need for distance-based and gradient-based algorithms. Standardization, implemented using by StandardScaler that is modifies features to possess a mean of 0 and a variance of one. this is mainly beneficial for algorithms like as Logistic Regression and Support Vector Machines. On the other hand, normalization achieved through MinMaxScaler that is rescales features to a prearranged range, typically between 0 and 1. this method is beneficial when features are measured in dissimilar elements or when algorithms are sensitive to differences in scale. Consequently, scaling promises that all features are presented on an analogous scale [21].

Standardization (Z-score Normalization):

$$z = \frac{x - \mu}{\sigma} \quad (4)$$

where (μ) is use for mean and (σ) is for standard deviation.

Min–Max Normalization:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (5)$$

Scales features into the range ([0,1]).

3.2.3 Categorical Encoding

Machine learning models needs to numerical inputs, so that categorical variables need to be encoded. Label Encoding, which allocates integer values to categories that is appropriate for ordinal features. In contrast, One-Hot Encoding makes binary columns for each category. This technique is frequently used for nominal variables, as it avoids the model from assuming any essential order among the categories [19].

$$\text{Category} \rightarrow 0,1,2,\ldots,k-1 \quad (6)$$

Each unique category is mapped to an integer.

3.2.4 Outlier Detection and Treatment

Outliers can twist model training and effect in the prediction accuracy. Predominant recognition techniques

incorporate statistical methods like as the Z-score and Interquartile Range (IQR) along with pictorial representations like boxplots. Outliers may be addressed by removal, overlaying or change, depending upon their influence and domain significance [21].

Z-score Method:

$$Z = \frac{x - \mu}{\sigma} \tag{7}$$

Values with (|Z| > 3) are naturally considered outliers.

Interquartile Range (IQR) Method:

$$IQR = Q_3 - Q_1 \tag{8}$$

Outliers if

$$x < Q_1 - 1.5 \times IQR; \text{or}; x > Q_3 + 1.5 \times IQR \tag{9}$$

A heart disease related patient showing no cholesterol level and no resting blood pressure is considered an outlier. Outliers are methodically extracted using the z-score of numerical columns in the dataset with a well-defined threshold for filtering. The dataset is divided into 80% of training data and 20% of testing data. MinMax Scaling is employed to standardize the features. Numerous baseline models were developed followed by 10-fold cross-validation and an ensemble approach was applied to classify the topmost performing baseline models. The models applied in this study comprise as Logistic Regression, Random Forest, Decision Tree, kNN, SVM and Naïve Bayes (NB). Model assessment is shown using performance metrics such as accuracy, precision, sensitivity, F1-score and correlation coefficient.

3.3 Machine learning models

This section describes a range of different types of ML algorithms are used, each selected for its separate ability to handle various kinds of data. The capability of ML models to analyse complicated datasets and discover significant understandings has made them a vital tool for HD prediction.

Forecasting the probability that a patient will develop cardiovascular disease (CVD) using clinical and demographic variables like as age, gender, BP levels, cholesterol levels, glucose, BMI and other lifestyle factors is a binary classification problem. In this feature vector as input

$$X = (x_1, x_2, \dots, x_n) \tag{10}$$

the task is to expect the class label

$$y \in 0,1 \tag{11}$$

where 1 denotes the presence and 0 denotes the absence of CVD.

3.3.1 Logistic Regression

Logistic Regression (LR) is employed as a baseline probabilistic model meanwhile it is simply interpretable and clinically applicable. Its usages the logistic function to calculate the probability of CVD incidence.

$$P(y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)}} \tag{12}$$

In this framework, (β_0) denotes the intercept and (β_i) denotes the coefficients of the model. LR is exclusively significant in CVD researches because the derived coefficients clearly demonstrate the effect of specific risk factors on the incidence of the disease [19], [20].

3.3.2 Decision Tree

Decision Trees (DT) classify patients by recursively partitioning the feature space through the application of decision rules that are derived from impurity measures, including the Gini Index.

$$\text{Gini} = 1 - \sum_{k=1}^{K} p_k^2 \tag{13}$$

Let p_k represent the proportion of samples that fit in the class k. Decision Trees are appropriate for CVD diagnosis because of their rule-based interpretability, which corresponds with clinical decision-making processes [21].

3.3.3 Random Forest

Random Forest (RF) is an ensemble learning technique that combines several decision trees, each trained on bootstrapped samples of the dataset. The ultimate prediction is derived from the process of majority voting.

$$\hat{y} = \text{mode}\{h_1(X), h_2(X), \dots, h_T(X)\} \tag{14}$$

where (h_t)denotes the prediction of the (t^{th})decision tree. Random Forest (RF) adeptly identifies non-linear relationships among cardiovascular risk factors and mitigates overfitting, hence enhancing predictive accuracy in cardiovascular disease (CVD) datasets [22], [23].

3.3.4 k-Nearest Neighbors

The k-Nearest Neighbors (k-NN) algorithm assesses cardiovascular disease (CVD) risk by determining the (k) most analogous patient records through a distance metric, such as Euclidean distance.

$$d(X_i, X_j) = \sqrt{\sum_{m=1}^{n} (x_{im} - x_{jm})^2} \tag{15}$$

The class label is resolute through majority voting between the closest neighbours. k-NN demonstrates to be helpful for predicting cardiovascular disease when the similarity among patients is a critical factor, its efficiency can be influenced by the scaling of features and the size of the dataset [24].

3.3.5 Naive Bayes

Naive Bayes (NB) is a probabilistic classifier that is constructed on Bayes' theorem [25]:

$$P(C \mid X) = \frac{P(X \mid C)P(C)}{P(X)} \quad (16)$$

Under the uncertain independence assumption, the probability is decomposed as:

$$P(X \mid C) = \prod_{i=1}^{n} P(x_i \mid C) \quad (17)$$

3.3.6 Support Vector Machine

Support Vector Machine (SVM) concepts an optimal hyperplane that make best use of the margin between classes:

$$\min_{w,b} \quad \frac{1}{2}|w|^2 \quad \text{s.t.} \quad y_i(w \cdot x_i + b) \geq 1 \quad (18)$$

For non-linear cardiovascular disease data, kernel functions such as the Radial Basis Function (RBF) are employed:

$$K(x_i, x_j) = \exp\left(-\gamma |x_i - x_j|^2\right) \quad (19)$$

Support Vector Machines have revealed remarkable generalization abilities in predicting cardiovascular diseases, specifically within high-dimensional feature spaces [26], [27].

3.4 Feature selection

Choosing the right features to attain optimal outcomes in data classification has emerged as a significant challenge in recent decades. While incorporating additional features can enhance prediction accuracy from a theoretical standpoint, practical evidence suggests that this is not universally applicable, as not every feature is critical for identifying the data class label. Certain elements do not pertain to the data level. Feature selection policies can be classified into three types like as filtering, wrapper and embedded [28].

3.4.1 Filtering methods

Filtering techniques evaluate the precision of predictions or categorizations using a non-direct metric, like the distance metric, which shows the degree of separation between the classes. As a preprocessing step, this approach is commonly employed. Actually, characteristics are chosen according to how well they do on several statistical tests that are used to link them to the outcome variable [29].

3.4.2 Wrapper methods

Wrapper methods employ a search algorithm in conjunction with a learning model to assess a subset of inheritable factor during the search process. Utilizing a learning model, wrapper methods generally provide superior classification achievement compared to filter methods. Conversely, they present several drawbacks, including substantial computational burden and the potential for overfitting [28, 29].

3.4.3 Embedded methods

These approaches are commonly given to learners and carry out feature selection during the learning process. By employing distinct evaluation criteria at each stage of the search process, this model builds upon earlier models as well. The qualities of filters and wrapper methods are combined in embedded methods. Internal feature selection approaches are used by algorithms to accomplish this. The common methods are LASSO and Decision tree [28-30].

3.5 Ensemble Model Development

The selected features are make used to train numerous base classifiers to establish an ensemble learning framework. Let $h_1(x), h_2(x), \ldots, h_M(x)$ denote M individual classifiers.

3.5.1 Base Learners

The ensemble participates a diversity of classifiers like as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbors and Naive Bayes to progress model diversity and minimize generalization error [29].

3.5.2 Voting-Based Ensemble Strategy

A Voting Classifier is implemented to combine predictions from several base learners.

Hard Voting:

$$\hat{y} = arg \ \max_{c} \sum_{m=1}^{M} \mathbb{I}(h_m(x) = c) \quad (20)$$

Soft Voting:

$$\hat{y} = \arg\max_{c} \sum_{m=1}^{M} P_m(y = c \mid x) \quad (21)$$

Soft voting is beneficial when probabilistic outputs are available as it takes into account the self-confidence levels of each model.

3.5.3 Model Training and Validation

Stratified sampling is make used to split the dataset into training and testing sets while conserving the class distribution. In order to maximize the hyperparameters tuning the models are trained on the training set and then

cross-validated. The last prediction from the ensemble is given by [30]:

$$H(x) = \sum_{m=1}^{M} w_m \, h_m(x) \text{subject to} \sum_{m=1}^{M} w_m = 1 \quad (22)$$

where w_m denotes the weight allocated to respectively base classifier.

3.6 Performance Evaluation

The effectiveness of the proposed ensemble model is measured by utilizing predictable classification metrics such as Accuracy, Precision, Recall, F1-score and Area Under the ROC Curve (ROC-AUC). These metrics offer a systematic estimation of the model's prediction efficiency and clinical consequence [28].

A predictive model is estimated according to four parameters that are true negative (TN) indicates accurate predictions for individuals without HD, true positive (TP) denotes accurate predictions for patients with HD, false positive (FP) specifies inaccurate classification of patients without HD as having HD and false negative (FN) replicates the incorrect classification of patients with HD as healthy. The assessment of the executed system's performance utilized the following metrics:

Accuracy: measures the overall complete correctness of predictions and formula is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (23)$$

Precision: enumerates the amount of correctly predicted positive instances among all predicted positives:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (24)$$

Recall: evaluates the model's capacity to correctly classify actual positive cases, which is mainly significant in clinical risk prediction:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (25)$$

F1-score: defined as the sympathetic mean of Precision and Recall, stabilities false positives and false negatives:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (26)$$

Additionally, the Area Under the Receiver Operating Characteristic Curve (AUC–ROC) assesses the model's capacity to differentiate between classes at several threshold levels:

$$\text{AUC} = \int_0^1 TPR(FPR)\, d(FPR) \quad (27)$$

where TPR symbolizes the true positive rate and FPR symbolizes the false positive rate.

Algorithm: Proposed Ensemble Framework

Input : Dataset D = {(xi , yi)}i=1N

Set of base classifiers H = {h1, h2, … , hM}

Number of selected features k

Output: Final ensemble prediction ŷ

Step1: Data Preprocessing

Handle missing values in D using mean/median (numerical) and mode (categorical)

Scale numerical features using StandardScaler or Min–Max normalization

Encode categorical features using Label Encoding or One-Hot Encoding

Detect and treat outliers using IQR or Z-score method

Step2: Feature Selection using Chi-Square

for each feature fj ∈ D do

Compute χ^2(fj , y)

end for

Rank all features based on χ^2 scores

Select top k features to obtain reduced dataset D′

Step3: Model Training

Split D′ into training set Dtrain and testing set Dtest

for each classifier hm ∈ H do

Train hm using Dtrain

Obtain predictions/probabilities on Dtest

end for

Step4: Ensemble Prediction

if Hard Voting then

ŷ ← argmaxc Σm I(hm(x) = c)

else if Soft Voting then

ŷ ← argmaxc Σm Pm(y = c | x)

end if

Step5: Performance Evaluation

Compute Accuracy, Precision, Recall, F1-score, and AUC

return ŷ

4 Result:

Feature extraction is a key strategy for dipping dimensionality in data preprocessing. This is because datasets often have redundant and irrelevant qualities. These variables can severely affect the effectiveness and complexness of classification algorithms.

Feature mining has two primary goals: to reduce the number of characteristics and to increase classification efficiency, which is a natural result of the process [31].

This study used the system to diagnose cardiovascular disease shown in Figure 3. The system used the chi-square feature selection method and a voting ensemble model. This ensemble model combined four different machine learning models as Logistic Regression, Decision Tree, Random Forest, SVC, K-Nearest Neighbors, and Naive-Bayes. using the chi-square feature selection algorithm, the dataset's twelve features were ranked by their importance, as shown in Figure 5.

The most important variables for diagnosing heart disease were identified as: ST slope, oldpeak, exercise angina, maximum heart rate, fasting blood sugar, cholesterol levels, the type of chest pain, and sex. using the chi-square method, we condensed the number of diagnostic features from twelve to eight, which helped to lessen the computational load.

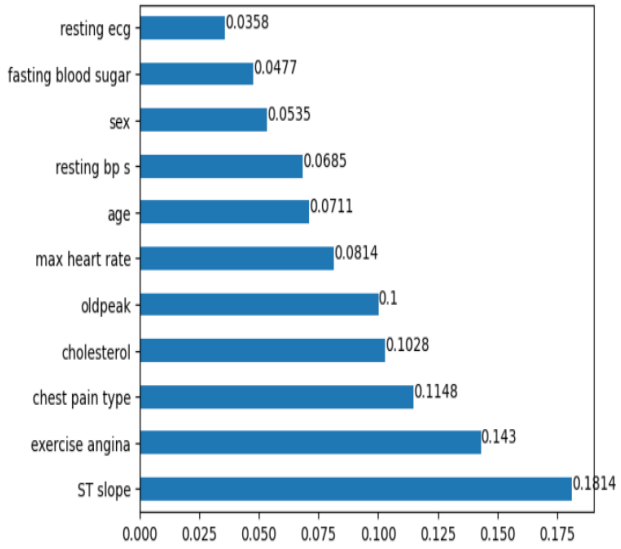

Figure 5: Features importance in prediction, according to their chi-square

Figure 6 illustrates a heat map that represents a correlation matrix of the features within the HD dataset. This representation pictorial represents the correlations between feature pairs with colour strength indicating both the direction and correlation coefficients. This heat map is an important tool for examining and understanding the complex relations among several features in the dataset which inform the following data preprocessing and modelling phases.

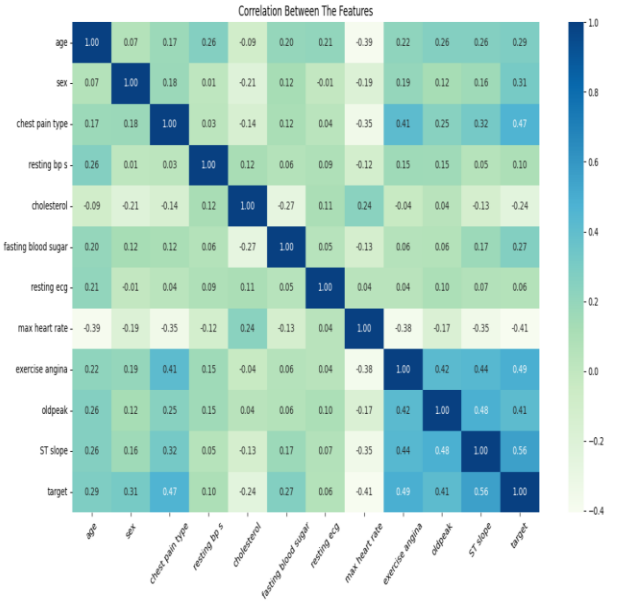

Figure 6: Correlation matrix between the features

In the assessment, two methodologies were studied. The preliminary method consisted of standardizing the HD dataset which involved 12 input features followed by the direct training and calculation of the individual models like as Logistic Regression, Decision Tree, Random Forest, SVC, K-Nearest Neighbors, Naive-Bayes without the application of the chi-square algorithm. The prediction outcomes from the base models were subsequently combined into the voting ensemble classifier. Once the base model has shaped its predictions then the ensemble classifier assesses them for inaccuracies before expressing its own prediction. This improvement allowed the cardiovascular diagnosis system to progress its complete accuracy while simultaneously reducing the error shaped by each specific base classifier. Table 1 and Figure 7 illustrate that the preliminary approach produced strong performance from the Random Forest and K-Nearest Neighbors classifiers, achieving the highest accuracy of 88.95% and 88.40% respectively, equated to the other four base classifiers. The performance of SVC and Decision Tree was particularly inferior than that of the other base classifiers, achieving accuracies of 87.85% and 85.08%, respectively. On the other hand, the voting ensemble classifier demonstrated superior performance compared with the base classifiers, accomplishing an accuracy of 89.50%, which replicates a significant 1.10% enrichment in accuracy for the top-performing base classifier (K-Nearest Neighbors).

Table 1: Performance metrics of models before feature selection

Model	Features	Accuracy	F1 Score	Precision	Recall
Logistic Regression	12	84.53%	85.57%	84.69%	86.46%
Decision Tree	12	85.08%	86.29%	84.16%	88.54%
Random Forest	12	88.95%	89.80%	88.00%	91.67%
SVC	12	87.85%	88.78%	87.00%	90.62%
K-Nearest Neighbors	12	88.40%	89.12%	88.66%	89.58%
Naive-Bayes	12	85.08%	85.71%	87.10%	84.38%
Proposed Ensemble Voting	12	89.50%	90.55%	89.22%	91.92%

Table 2: Performance metrics of models after feature selection

Model	Features	Accuracy	F1 Score	Precision	Recall
Logistic Regression	8	87.29%	88.56%	87.25%	89.90%
Decision Tree	8	83.98%	84.97%	87.23%	82.83%
Random Forest	8	88.95%	90.00%	89.11%	90.91%
SVC	8	87.85%	89.22%	86.67%	91.92%
K-Nearest Neighbors	8	86.19%	87.05%	89.36%	84.85%
Naive-Bayes	8	85.08%	86.43%	86.00%	86.87%
Proposed Ensemble Voting	8	90.61%	91.46%	91.00%	91.92%

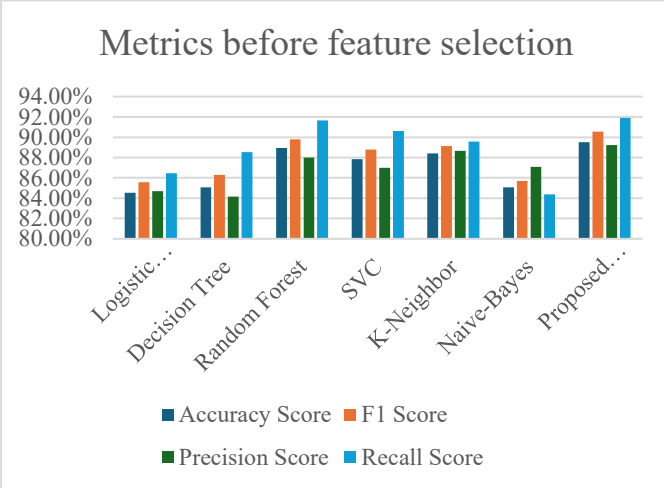

Figure 7: Models Performance before feature selection

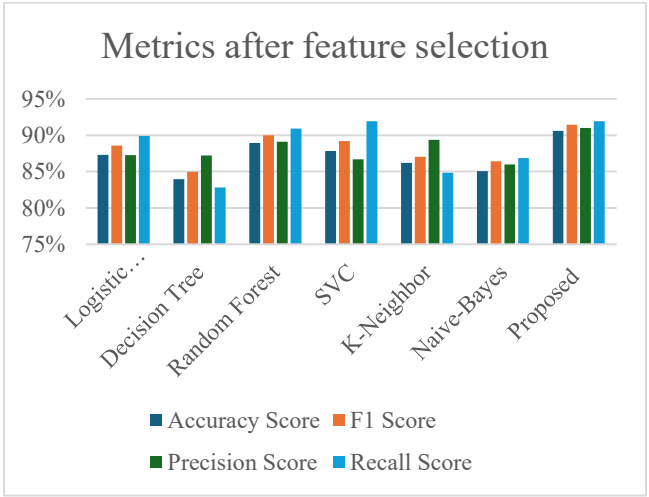

Figure 8: Models Performance after feature selection

In the second approach, the cardiovascular disease dataset was normalized by using feature scaling technique, followed by the selection of eight main features employing by the chi-square feature selection technique. The individual models like as Logistic Regression, Decision Tree, Random Forest, SVC, K-Nearest Neighbors, Naive-Bayes were subsequently trained and evaluated utilizing the concentrated-on feature dataset. The prediction outcomes from the base models were later input into the voting ensemble model then the resulting in the final prediction. Table 2 and Figure 8 represent the accuracy of individually base classifier as well as the ensemble voting model ensuing feature reduction. The findings specified an enhancement in achievement for the RF, SVC and LR models with feature selection while the DT and NV models displayed no enhancement. The accuracy scores for the RF, SVC and LR models were 88.95%, 87.85%, and 87.29% respectively. The voting ensemble classifier exceeded the base models that is attaining an accuracy of 90.61%, which signifies a significant 1.66% improvement over the highest performing base classifier i.e. Random Forest. The use of the chi-square feature selection technique in combination with the projected voting ensemble classifier improves the problem-solving capabilities for CVD.

Figure 9 represents a comparative analysis of the prediction accuracy accomplished by the two different methodologies. The figure determines that the Random Forest (RF) and K-Nearest Neighbors (KNN) models yield comparable performance levels that is achieving the maximum accuracy between the four initial classifiers. On the other hand, the Support Vector Classifier (SVC) and Decision Tree (DT) models show reduced accuracy across both the approaches. Additionally, the voting classifier's higher performance as specified in the figure, suggests that ensemble techniques successfully capitalize on the strengths essential in various models by this means enhancing overall performance. The figure also discloses that the Feature Selection (FS) technique substantially progresses the accuracy of the base models and accordingly the accuracy of the voting ensemble concluding in an ultimate accuracy rate of 90.61%.

This represents a 1.25% enhancement over the voting ensemble model that did not employ the Feature Selecting methodology. Additionally, the precision, recall and F1-score metrics for the six different classifiers along with the voting ensemble classifier were intended both previous to and subsequent to the Feature Selection technique as illustrated in Figures 10, 11 and 12. These figures determine that the proposed voting ensemble model

reliably betters the performance of the individual classifiers transversely all evaluated metrics, regardless of whether the Feature Selection method was implemented.

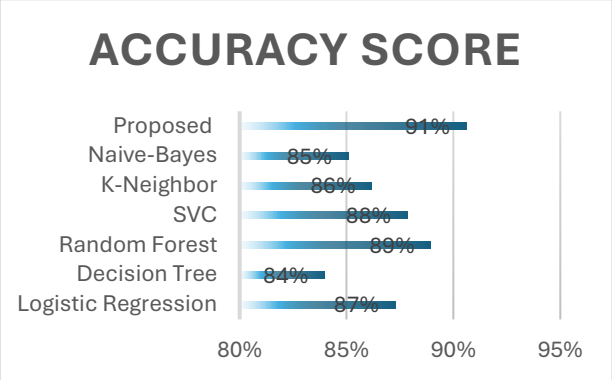

Figure 9: Accuracy Score

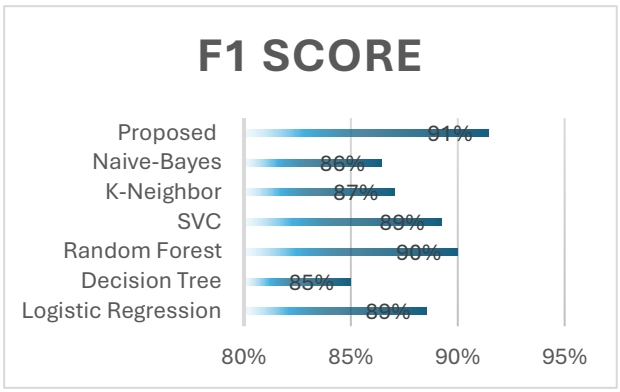

Figure 10: F1-Score

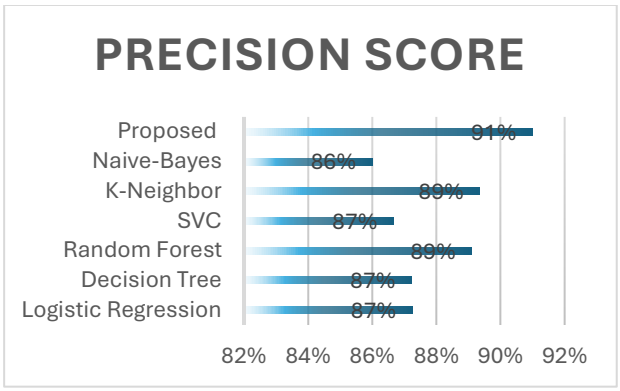

Figure 11: Precision Score

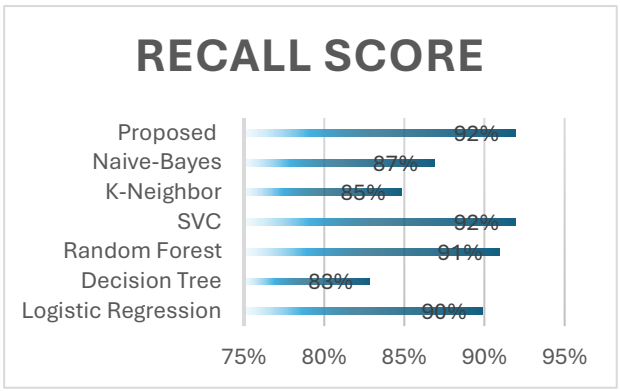

Figure 12: Recall Score

The observation of the confusion matrix progresses the accuracy results by demonstrating the classification behaviour of all models i.e. represented in Figure 13. The Decision Tree model displays a superior incidence of misclassifications, specifically false positives and false negatives which excess for its reduced accuracy. Naive Bayes and K-Nearest Neighbors demonstration moderate improvement. however, they remain to display significant classification errors. Logistic Regression and SVC determine a more reasonable distribution of true positives and true negatives which matches to their raised-up accuracy metrics. The Random Forest model improves classification accuracy by increasing the accurate predictions of both disease and no-disease cases. The proposed ensemble voting model provide the most valuable confusion matrix that is characterized by the maximum count of precisely classified instances and nominal errors. Ensemble learning improves the consistency and robustness of classification.

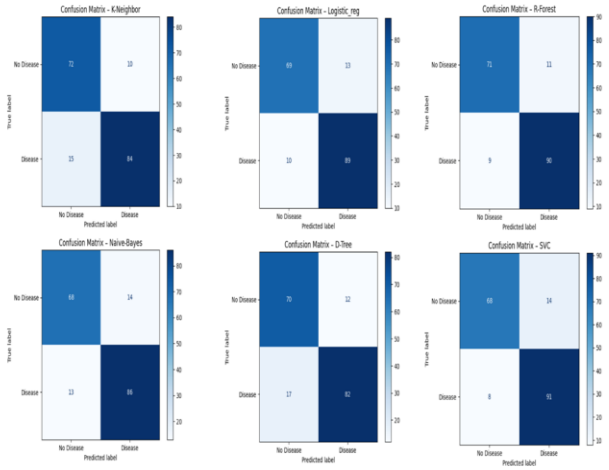

Figure 13: Confusion Matrix analysis of all the models

Additionally, to simplify a more detailed assessment of the indicative model's performance ROC and AUC analyses were employed. The ROC–AUC plot represents the false positive rate along with the x-axis and the true positive rate along with the y-axis. This plot assesses the model's measurements to discriminate amid two classes with 0 representing the absence and 1 representing its presence of disease. The perfect ROC curve is positioned in the upper left quadrant of the plot [28]. Therefore, an advanced ROC curve indicates that the model is generating precise predictions and effectively distinguishing between the two classes.

A model's capability to distinct classes is strong when the AUC is close to one (1) and weak when the AUC is close to zero (0). An AUC of 0.5 recommends the model can't differentiate between classes [31]. We used the ROC–AUC curve to evaluate how well fit the base models and the voting ensemble predicted cardiovascular disease as demonstrated in Figure 14. The figure demonstrations that

the AUC scores for all classifiers reduced after feature selection (FS). This recommends that feature selection had slight impact on the AUC curve.

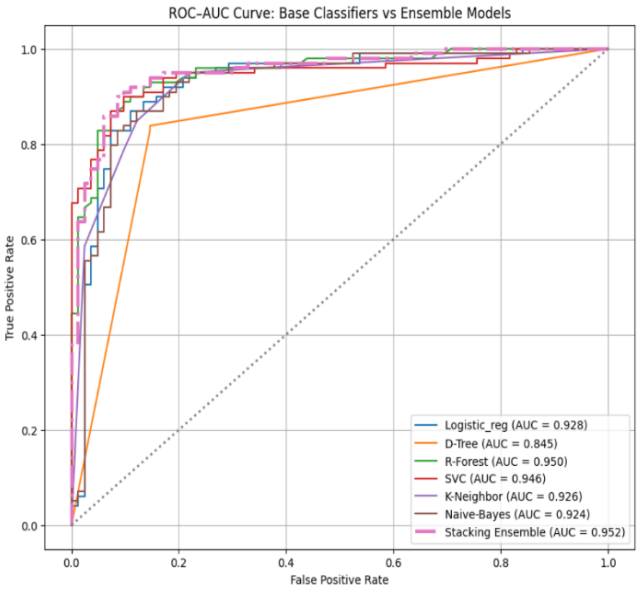

Figure 14: ROC-AUC chart for models

5. Discussion:

This paper improves cardiovascular information processing by introducing a more precise, robust and understandable cardiovascular disease risk prediction framework that integrates the advantages of ensemble learning and statistically based feature selection. The recommended model tackles important issues in recent risk prediction methodologies including unnecessary dimensionality, data noise and limited interpretability. The system rallies indicative accuracy and facilitates the early detection of high-risk individuals by including Chi-Square feature selection and enhanced ensemble classifiers. This directly effects the enhancement of pre-emptive treatment and simplifies greater clinical decision-making and may decrease the global burden of cardiovascular diseases.

6. Conclusion:

The combined Cardiovascular Disease Dataset helped as the foundation for this study. Preprocessing involved the amputation of outliers which is achieved through the application of the interquartile range (IQR) and tuning. which was accomplished by using Min-Max normalization techniques.

This research represents a cardiovascular disease detection system predicated on a voting ensemble method and a chi-square feature selection technique. The ensemble model combined a variety of base machine learning techniques exactly like as Logistic Regression, Decision Tree, Random Forest, SVC, K-Nearest Neighbors and Naive-Bayes. The heart_statlog_cleveland_hungary_final cardiovascular ailment dataset was employed for the training and testing phases of both the ensemble and base models.

The model's effectiveness was measured through a variety of metrics about accuracy, precision, recall, the confusion matrix, the F1-score and the area under the curve (ROC-AUC). The introductory classifiers specifically as Logistic Regression, Decision Tree, Random Forest, SVC, K-Nearest Neighbors and Naive-Bayes unveiled accuracies of 84.53%, 85.08%, 88.95%, 87.85%, 88.40%, and 85.08%, respectively. In contrast, the voting ensemble classifier outperformed the base classifiers achieving a higher accuracy of 89.50%. Also, the application of the chi-square feature selection method revealed eight significant features which subsequently improved performance. thus, increasing the voting ensemble model's accuracy to a notable 90.61%.

7. Limitations:

• The Chi-Square method only evaluates non-negative and categorical features, requiring discretization of continuous variables, which may affect information retention.

• The study does not incorporate advanced deep learning architectures or multimodal data sources (e.g., images, ECG signals).

• The model's generalizability may be constrained by the characteristics of the chosen dataset and may require validation on external clinical data.

• Clinical interpretability remains limited to ensemble-based feature importance and does not provide causal explanations.

8. Future work:

The flexibility and scalability of the framework will be assessed in future studies by confirming it using large-scale, real-world medical data. Wearable sensor data when united with data from non-invasive sensors may also improve early detection. Finally, while this study did use standard machine learning models, future research could aspect into integrating with deep learning to improve feature learning and prediction accuracy in complicated, high-dimensional datasets. Taken as a whole, these recommendations should make the proposed approach more flexible and valuable in a variety of healthcare situations.

References:

1. WHO. Cardiovascular diseases (CVDs), 11 June 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). [Accessed 15 October 2025].

2. GBD 2023 Study, "Global Burden of Cardiovascular Diseases, 1990–2023," The Lancet, 2024.

3. D'Agostino et al., "General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study," Circulation, 2008.

4. Recent advances in ML for CVD prediction (e.g., arXiv:2401.17328, 2024).

5. Ay, Ş., Ekinci, E. & Garip, Z. A comparative analysis of meta-heuristic optimization algorithms for feature

selection on ML-based classification of heart-related diseases. J. Supercomput. 1–30 (2023).

6. Zaini, N. A. M. & Awang, M. K. Hybrid feature selection algorithm and ensemble stacking for heart disease prediction. Int. J. Adv. Comput. Sci. Appl. https://doi. org/ 10. 14569/ IJACSA. 2023. 01402 20 (2023).

7. Mienye ID, Sun Y, Wang Z. An improved ensemble learning approach for the prediction of heart disease risk. Inform Med Unlocked 2020; 20: 100402. http://dx.doi.org/10.1016/j.imu.2020.100402

8. Spencer R, Thabtah F, Abdelhamid N, Thompson M. Exploring feature selection and classification methods for predicting heart disease. Digit Health 2020; 6: 2055207620914777. http://dx.doi.org/10.1177/2055207620914777 PMID: 32284873

9. Aziz S, Aslam Z, Rizwan M, Nawaz S. EArly heart disease prediction with minimal attributes using machine learning. Pak J Eng Technol 2020; 3(2): 178-82. http://dx.doi.org/10.51846/vol3iss2pp178-182

10. Day JR, Gupta A, Abro C, Jung K, Krishnamurti L, Takemoto C, et al. Risk factors and cardiovascular disease (CVD) related outcomes in hospitalized patients with hemophilia 10-year follow-up. Blood. (2020) 136:30–1. doi: 10.1182/blood 2020-136486

11. Li JP, Haq AU, Din SU, Khan J, Khan A, Saboor A. Heart disease identification method using machine learning classification in e healthcare. IEEE Access 2020; 8: 107562-82. http://dx.doi.org/10.1109/ACCESS.2020.3001149

12. Chen, H.L., Yang, B., Liu, J., Liu, D.Y.: A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Expert Syst. Appl. 38(7), 9014–9022(2011).https://doi.org/10.1016/j.eswa.2011.01.120 13.

13. Wang, S.-J., Mathew, A., Chen, Y., Xi, L.-F., Ma, L., Lee, J.: Empirical analysis of support vector machine ensemble classifiers. Expert Syst. Appl. 36(3), 6466–6476 (2009). https://doi.org/10. 1016/j.eswa.2008.07.041 14.

14. Kahramanli, H., Allahverdi, N.: Design of a hybrid system for the diabetes and heart diseases. Expert Syst. Appl. 35(1–2), 82–89 (2008). https://doi.org/10.1016/j.eswa.2007.06.004 15.

15. Raza, K. Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule. in U-Healthcare Monitoring Systems, Design and Applications (eds Dey, N. et al.) (Academic, 2019).

16. Mienye, I. D., Sun, Y. & Wang, Z. An improved ensemble learning approach for the prediction of heart disease risk. Inf. Med. Unlocked. 20, 100402 (2020).

17. Ambrews, A. B. et al. Ensemble based machine learning model for heart disease prediction. in International Conference on Communications, Information, Electronic and Energy Systems (CIEES) (2022).

18.Dataset link https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive#9

19. D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed., Wiley, 2013.

20. P. K. Shah et al., "Risk prediction of cardiovascular disease using logistic regression," IEEE Access, vol. 8, pp. 123456–123465, 2020.

21. J. R. Quinlan, "Induction of decision trees," Machine Learning, vol. 1, no. 1, pp. 81–106, 1986.

22. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

23. A. K. Singh and S. Gupta, "Heart disease prediction using random forest," Procedia Computer Science, vol. 132, pp. 135–143, 2018.

24. T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Trans. Information Theory, vol. 13, no. 1, pp. 21–27, 1967.

25. G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," Proc. 11th Conf. Uncertainty in AI, pp. 338–345, 1995.

26. V. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.

27. S. Polat and H. D. Mehr, "Classification of cardiovascular disease using SVM," Expert Systems with Applications, vol. 38, no. 10, pp. 12347–12354, 2011.

28. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182 (2003)

29. Tuv, E., Borisov, A., Runger, G., Torkkola, K.: Feature selection with ensembles, artificial variables, and redundancy elimination. J. Mach. Learn. Res. 10, 1341–1366 (2009)

30. Vergara, J.R., Estévez, P.A.: A review of feature selection methods based on mutual information. Neural Comput. Appl. 24(1), 175–186 (2014). https://doi.org/10.1007/s00521-013-1368-0

31. Shah, F.P., Patel, V.: A review on feature selection and feature extraction for text classification. In: 2016 international conference on wireless communications, signal processing and networking (WiSPNET) (pp. 2264–2268), IEEE (2016). https://doi.org/10.1109/WiSPNET.2016.7566545