Journal of Carcinogenesis
Covering Genesis to Therapeutics

Journal of Carcinogenesis, ISSN: 1477-3163
2025; Vol 24: Issue 2s, https://carcinogenesis.com/

# Spline Regressive Quadratic Emphasis Boosted Ensemble Classifier For Heart Disease Prediction

**N.Haridoss [1], D. Ashok Kumar [2]**

[1]Research Scholar, Department of Computer Science, Government Arts & Science College (Affiliated to Bharathidasan University, Tiruchirappalli) Kumulur, Lalgudi-621601.

Email ID : harivdm@gmail.com

[2] Associate Professor and Head, Department of Computer Science, Government Arts & Science College (Affiliated to Bharathidasan University, Tiruchirappalli) Kumulur, Lalgudi-621601.

Email ID: drashoktrichy22@gmail.com

## ABSTRACT

Heart disease is a major global health concern affects the heart and blood arteries like arrhythmias, heart failure, and coronary artery disease. Heart diseases are the leading cause of the severe mortality rate worldwide for both men and women. Early detection of heart disease plays a vital role for timely treatment and continuous monitoring by healthcare providers, and reducing mortality rates. Numerous conventional machine learning methods are developed to identify the heart disease over the decades. However, these models faced the challenges of accurate prediction with minimal time consumption. In order to enhance the heart disease prediction accuracy, a novel method called Spline Regressive Quadratic Emphasis Boosting Classifier (SRQE-Boost) model is proposed. The main aim of proposed SRQE-Boost model is to perform the heart disease prediction through the significant feature selection and classification to minimize the time as well as the space consumption. The proposed SRQE-Boost model comprises four processes, namely data acquisition, preprocessing, feature selection and classification. The data acquisition process is the first step for predicting heart disease with large volume of patient data collected from the input dataset. After data acquisition, preprocessing is carried out to minimize the time as well as memory consumption. During data preprocessing, missing data handling using linear spline interpolation method and outlier removal based on Peirce criterion are carried out to organize the dataset into a suitable format. Followed by, feature selection process is employed using factor regressive analysis to select the relevant features to improve the heart disease prediction by minimizing the dimensionality of the dataset. Factor regressive analysis is a type of statistical analysis used for data analysis through measuring the relationships between features and the target based on Tanimoto Similarity Index. Finally, Quadratic Discriminant Emphasis Boosting ensemble classifier is employed for predicting the heart disease presence or absence with the selected features. In this way, accurate heart disease prediction results are observed with minimal time consumption. Experimental evaluation is carried out on performance metrics like accuracy, precision, recall, F1 score, specificity, AUC, MCC, Prediction time, memory consumption, with respect to number of data samples and features. Quantitative analysis results indicate that the proposed SRQE-Boost model achieved better accuracy in disease prediction, and minimizes time as well as memory consumption compared to existing methods.

***Keywords:*** *Heart disease prediction, linear spline interpolation method, Peirce criterion, factor regressive analysis, Tanimoto Similarity Index, Quadratic Discriminant Emphasis Boosting ensemble classifier.*

**How to Cite:** N.Haridoss , D. Ashok Kumar , (2025) Spline Regressive Quadratic Emphasis Boosted Ensemble Classifier For Heart Disease Prediction, *Journal of Carcinogenesis*, *Vol.24, No.2s*, 248-276

## 1. INTRODUCTION

Heart disease, also known as cardiovascular disease (CVD), includes a range of conditions affecting the heart and blood vessels, heart failure, arrhythmias, and it leading cause of death worldwide. Reliable and early forecasting of heart disease is crucial for efficient patient management. Accurately heart disease prediction is crucial for determining efficient cardiac treatment options as the volume of data grows exponentially. Application of different machine learning algorithms has been developed for enhancing the results in predicting the risk of heart disease, to improve clinical decision-making.

A Logistic regression (LR) machine learning model was developed in [1] based on Boruta feature selection model with the aim of accurately detecting the heart disease. However, the method failed to apply on a diverse large volume of heart disease dataset with more instances and attributes. A stacking-based classification model [2] with firefly optimization algorithm was designed to enhance the accuracy of heart disease prediction. However, the designed algorithm failed to

find the global optimum feature.

A quantum-enhanced machine learning model was introduced in [3] for heart disease prediction. The designed model reduces the training time but it experiences errors during the learning process. Two supervised learning-based machine learning techniques were introduced in [4] for prediction of heart disease. But the model failed to enhance the performance of precision during heart disease prediction. An ensemble learning algorithms integrated with explainable AI model was developed in [5] to enhance the sensitivity and specificity in heart disease prediction. But designed model failed to enhance its disease predictive performance when applied to a large volume of patient data. A hybrid deep learning algorithm was designed in [6] for heart disease detection with large data analysis using Recursive feature elimination method. However, the recursive feature elimination model was more time consuming especially for large dataset. An advanced boosting ensemble technique was developed in [7] for the prediction of cardiovascular diseases using correlation feature-based selection. But the correlation-based feature was not effectively detecting the linear relationship between the features. A dual-stage stacked machine learning (ML) algorithm was designed in [8] for heart disease risk prediction. However, the efficient feature selection algorithms were not deployed for cardiac disorders prediction using the larger instances of a dataset. An improved explainable learning-based technique was designed in [9] for heart disease prediction through the integration of data normalization and feature selection. However, the precision evaluation performance was not addressed

A hybrid deep learning model was introduced in [10] for coronary heart disease prediction using discriminative features extraction with minimal computation cost. However, it failed to use efficient computational techniques to improve the model performance to accurately and effectively prevent the heart disease. A quantum-behaved particle swarm optimization (QPSO) algorithm was designed in [11] to determine the optimal feature for heart disease risk detection by transforming nominal data into numerical data and applying effective scaling techniques. However, it did not focus on selecting the more efficient features for accurately predicting the heart disease severity level. The machine learning algorithm was designed in [12] for more accurate heart disease predictions. However, the performance of time complexity in heart disease risk prediction was high. Machine learning algorithms were developed in [13] to improve the performance of timely accurate heart disease diagnosis. However, it did not apply the efficient feature selection algorithms to reduce the complexity of the heart risk prediction. A novel artificial neural network (ANN) was developed in [14] for heart disease detection using features extraction. However, it failed to improve accuracy by enhancing sensor performance and sensitivity. A new self-attention-based transformer model was introduced in [15] to enhance cardiovascular risk prediction by effectively modeling complex data. However, the model failed to improve its performance, especially when dealing with limited labeled data.

### 1.1 Key contributions

This section revolves key contributions SRQE-Boost model are outlined below:

To enhance the heart disease prediction, a novel SRQE-Boost model has been proposed by integrating data preprocessing, feature selection and classification.

To minimize the feature selection time, a SRQE-Boost model method has performs data pre-processing and feature selection. The preprocessing step includes missing data handling and outlier data removal using spline interpolation method and Peirce criterion method respectively. Tanimoto indexive factor regressive analysis is employed to select the relevant features and remove the other irrelevant features to improve the heart disease prediction by measuring the relationships between features and the target. This process also reduces the memory consumption

To improve the accuracy of disease prediction, Quadratic Discriminant Emphasis Boosting ensemble classifier is employed with a set of optimal features. The stacking method provides accurate classification output and minimizes the error.

Finally, a comprehensive evaluation is carried out to assess the performance of the heart disease prediction using various metrics and comparing it to other classification methods.

### Organization

This paper is organized into different sections as follows: Section 2 provides a review of related works in heart disease prediction. Section 3 introduces the proposed SRQE-Boost model, including a detailed explanation with clear diagram. Section 4 describes experimental settings and dataset description. Section 5 evaluates the performance of the proposed SRQE-Boost model in comparison to existing methods using various metrics. Finally, Section 6 presents the conclusions of the paper.

## 2. RELATED WORKS

A hybrid model was developed in [16] to enhance model accuracy for effectively detecting cardiovascular disease by selecting the highly relevant features. However, it failed to focus on applying a large datasets to enhance the efficiency of the model. A scalable machine learning-algorithm was developed in [17] for early cardiovascular disease prediction based optimal feature selection. The Fast Correlation-Based Filter Solution (FCBF) was employed from large-scale datasets for identifying relevant features and improving the performance of algorithms. Though the FCBF model increases the accuracy, time efficient features selection was major issues. A chi-square based feature selection was developed in [18]

with the aim of cardiovascular disease detection. But the designed algorithm provided the insignificant results when dealing with large samples. A multidimensional feature engineering and machine learning models were developed in [19] for accurate heart disease prediction. However, it did not developing the advancement and innovation of medical data analysis technology. An Explainable Artificial Intelligence (XAI) model was developed in [20] for heart disease classification. However, it failed to improve the early detection and personalized treatment strategies for heart disease.

An integration of artificial flora optimization algorithm with the SVM algorithm was developed in [21] for efficiently identifying the most significant features for heart disease prediction. However, it failed to include a data from diverse populations and incorporating a broader range of clinical and demographic features to improve the performance. A light gradient-boosting machine algorithm was designed in [22] to enhance its performance and accuracy of the heart disease prediction. However, it failed to perform the risk factors analysis. An explainable machine learning approach was designed in [23] to predict heart diseases. However, the model failed to apply the efficient preprocessing models for enhancing the heart disease prediction performance.

The whale optimization algorithm was introduced in [24] for feature selection and heart disease prediction. But, the designed algorithm faced the issues relating to high dimensional dataset. Machine learning techniques were developed in [25] with the aim of detecting the early detection of heart diseases. However, the designed technique failed to enhance the robustness of the models. A fuzzy logic based expert system was designed in [26] for the prediction and diagnosis of heart disease. However, the fuzzy logic system introduced high complexity in heart disease prediction. A hybrid feature selection model was developed in [27] for effective classification of cardiovascular disease. The models failed to focus on selecting the most discriminative attributes for disease prediction. A new hybrid ensemble learning approach was introduced in [28] that integrate multiple machine learning classifiers for heart disease prediction. However, the designed model failed to improve the accuracy of feature selection. A quantum machine learning model was developed in [29] to perform multi-class classification of cardiovascular diseases. However, early detection and prediction remained major challenges. An explainable AI based new deep learning model was developed in [30] for accurate heart disease prediction using Principal Component Analysis to reduce dimensionality, enhancing model efficiency. However, the time consumption of heart disease prediction was not reduced.

## 3. PROPOSED METHODOLOGY

This section describes the proposed methodology aimed at achieving heart disease prediction with minimal time consumption. To achieve this objective, a novel SRQE-Boost model is developed on a dataset to generate results with higher accuracy. To enhance the methodology, the input dataset needs to be cleaned, irrelevant information eliminated, and significant features selected. The improved methodology produces more accurate heart disease prediction results and superior model performance, as shown in Figure 1.

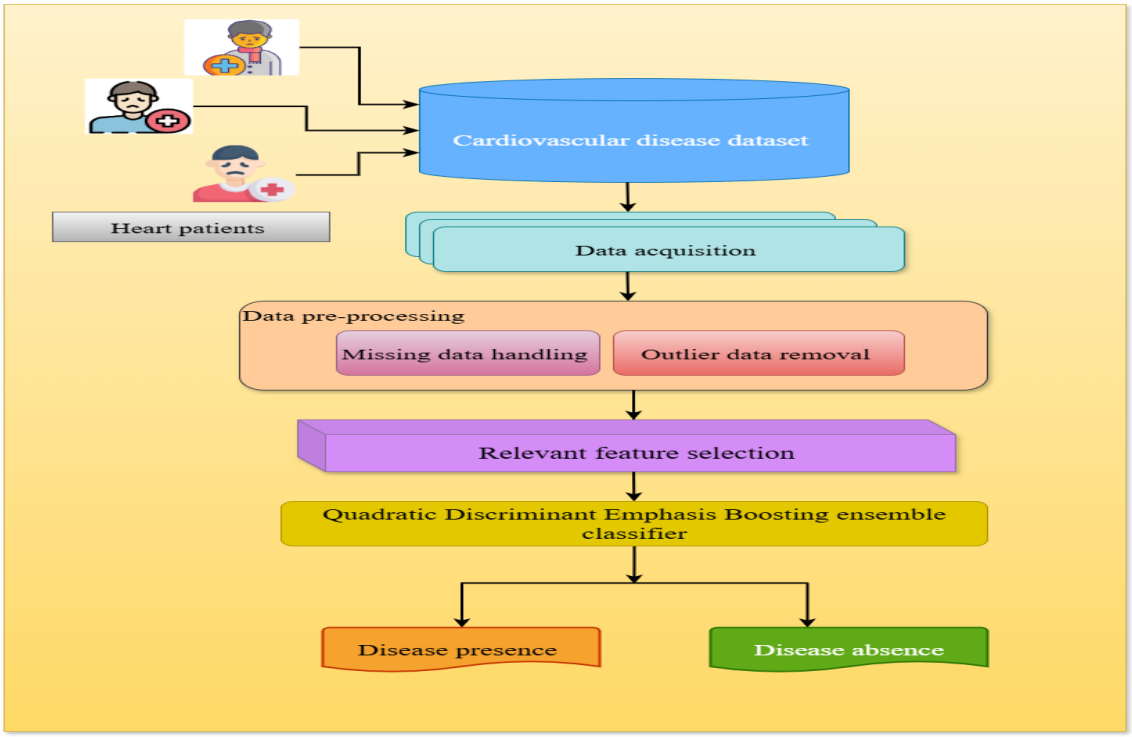



**Figure 1 Architecture Diagram of the SRQE-Boost model**

Figure 1 above portrays the architecture diagram of the proposed SRQE-Boost model for accurate heart disease prediction. The proposed SRQE-Boost model includes four fundamental processes namely data acquisition, preprocessing, feature selection and classification. These four fundamental processes are integrated to further enhance the accuracy of heart disease prediction with minimal time consumption. Therefore, the integration process of proposed SRQE-Boost model is explained briefly in the following subsections.

### 3.1 data acquisition

In the proposed SRQE-Boost model, data acquisition is the fundamental process of gathering the data using cardiovascular disease Dataset taken from https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset. By using these fundamental data acquisition step, the proposed method effectively acquire large volume of patient data for developing predictive models to perform the heart disease prediction. The dataset consists of 13 attributes and 70000 data samples. The attribute descriptions are listed in table 1.

**Table 1 Attribute Description**

| S.No | Attributes | Description |
|---|---|---|
| 1. | ID | Patient ID |
| 2. | Age | Patient age in days |
| 3. | Height | Patient height in cm |
| 4. | Weight | Patient weight in kg |
| 5. | Gender | 1-women, 2-men |
| 6. | ap_hi | Systolic blood pressure |
| 7. | ap_lo | Diastolic blood pressure |
| 8. | Cholesterol | Cholesterol<br>1: normal<br>2: above normal<br>3: well above normal |
| 9. | Gluc | Glucose<br>1: normal,<br>2: above normal<br>3: well above normal |
| 10. | Smoke | Smoking<br>1: Yes<br>0:no |
| 11. | Alco | Alcohol intake<br>1: Yes<br>0:no |
| 12. | Active | Physical activity |
| 13. | Cardio | 1 presence<br>0 absence |

Let us consider the cardiovascular disease Dataset '$DS$' comprises of patient data a sample '$DP$' as well as features $\{F_1, F_2, \dots, F_m\}$are organized in the form of matrix. Therefore, the input matrix with these dataset samples and features are formulated as follows,

$$M = \begin{bmatrix} F_1 & F_2 & \dots & F_m \\ PD_{11} & PD_{12} & \dots & PD_{1n} \\ PD_{21} & PD_{22} & \dots & PD_{2n} \\ \vdots & \vdots & \dots & \vdots \\ PD_{m1} & PD_{m2} & \dots & PD_{mn} \end{bmatrix} \quad (1)$$

Where, $M$ indicates an input matrix where each column indicates a number of features $F = \{F_1, F_2, \dots, F_m\}$, each row comprises of a number of data samples or patient data or instances '$PD = \{PD_1, PD_2, \dots, PD_n\}$' respectively.

The proposed SRQE-Boost model performs the data preprocessing tasks to organize the dataset before applying to machine learning. The data preprocessing step includes two major processes namely missing data as well as outliers' data within the input matrix.

The missing data refers to a no data value stored in cells for specific features within a dataset. This problem is handled by applying the linear spline interpolation method with other known data samples of particular features. A linear spline interpolation method is a method used to measure the values between known data points through piecewise linear functions. The proposed interpolation method simply connects adjacent data points with straight lines. These connected straight line is used to find the new missing data samples.

Let us consider the two adjacent data points coordinates' $(x_i, y_i)(x_{i+1}, y_{i+1})$ and the linear spline function is expressed as follows,

$$y = y_i + \frac{y_{i+1} - y_i}{x_{i+1} - x_i} . (x - x_i) \quad (2)$$

Where, $y$ denotes a missing data sample, $x$ is the location at which to estimate the missing value. From the above observed value, the missing data are handled in an accurate manner,

Followed by, the outlier's data detection and removal process is carried out from the dataset through the Peirce criterion. It is a statistical method used to determine the two or more outliers within the dataset. It is measured as the absolute difference between the data and their mean value is greater than the product of the maximum allowable deviation and standard deviation

$$Q = |PD_i - \mu| \quad (3)$$

$$PC = \begin{cases} Q > D_{max} * \sigma \ ; \ Outlier \\ Otherwise \ ; \ no\ outlier \end{cases} \quad (4)$$

Where, $PD_i$ denotes the data point, $\mu$ denotes a mean (or average) of the dataset, $D_{max}$ a ratio that defines the maximum allowable deviation from the mean, $\sigma$ denotes a standard deviation of the dataset. If the determined '$Q$' value is larger than the$D_{max} * \sigma$ , then the particular data is said to be a outlier. Otherwise the data is said to be a normal. The outlier data is removed and the missing data handling approach is employed to fill the data into respective cells. In this way, both missing data and outlier data handling processes are simultaneously performed. The processing algorithm is given below,

| // **Algorithm 1: Data pre-processing** |
|---|
| Input: cardiovascular disease dataset '$DS$', features $F = \{F_1, F_2, \dots, F_m\}$, patient data samples $PD = \{PD_1, PD_2, \dots, PD_n\}$ |
| **Output**: Pre-processed dataset |
| **Begin**<br>1. **For** each dataset '$DS$' with features '$F$' do<br>2. Formulate input vector matrix '$M$' using (1)<br>3. **If** any missing value in '$M$' **then**<br>**4.** **Apply** linear spline interpolation using (2)<br>5. Fill the missing value to the respective cell<br>6. **End if**<br>**7.** **For each data** samples with neighboring data samples **do**<br>**8.** Measure the difference between the mean and data samples using (3)<br>9. **if ( $Q > D_{max} * \sigma$) then**<br>10. Data samples is outlier |

```
11.    else
12.        Data samples is normal
13.    End if
14.    Remove outlier data
15.    Return (preprocessed dataset)
16.  End for
17. End for
 End
```

Algorithm 1 describes the heart disease dataset preprocessing to minimize the time as well as space consumption. Initially, a number of patient data are collected from the dataset and formulate the input matrix. Subsequently, missing data is identified and it filled by applying linear spline interpolation method. Once missing values are handled, the issue of outlier data removal is addressed. Firstly, the difference between the mean and the data samples are determined. If the estimated difference is lesser than the maximum allowable difference, then the sample is said to be a normal. Otherwise, it is said to be an outlier data. As a result, the preprocessed dataset is obtained at the output.

### 3.3 Tanimoto indexive factor regressive analysis based feature selection

With the preprocessed data set, the feature selection process is carried out with the aim of reducing the dimensionality of the dataset. Dimensionality reduction is a method to minimize the number of features and select the more related features within a big dataset. This process helps to minimize the computational complexity and challenges in achieving accurate heart disease prediction. To address this issue, the Tanimoto indexive factor regressive analysis is introduced in proposed SRQE-Boost model for dimensionality reduction by choosing the more related features. Through the identification of significant features, this approach helps to make the accurate prediction of heart disease.

Factor regression analysis model integrates factor analysis (minimizing features) and regression analysis (analyzing relationships). Factor regressive analysis is a type of machine learning technique used for data analysis through measuring the relationships between features and the target based on Tanimoto similarity index.

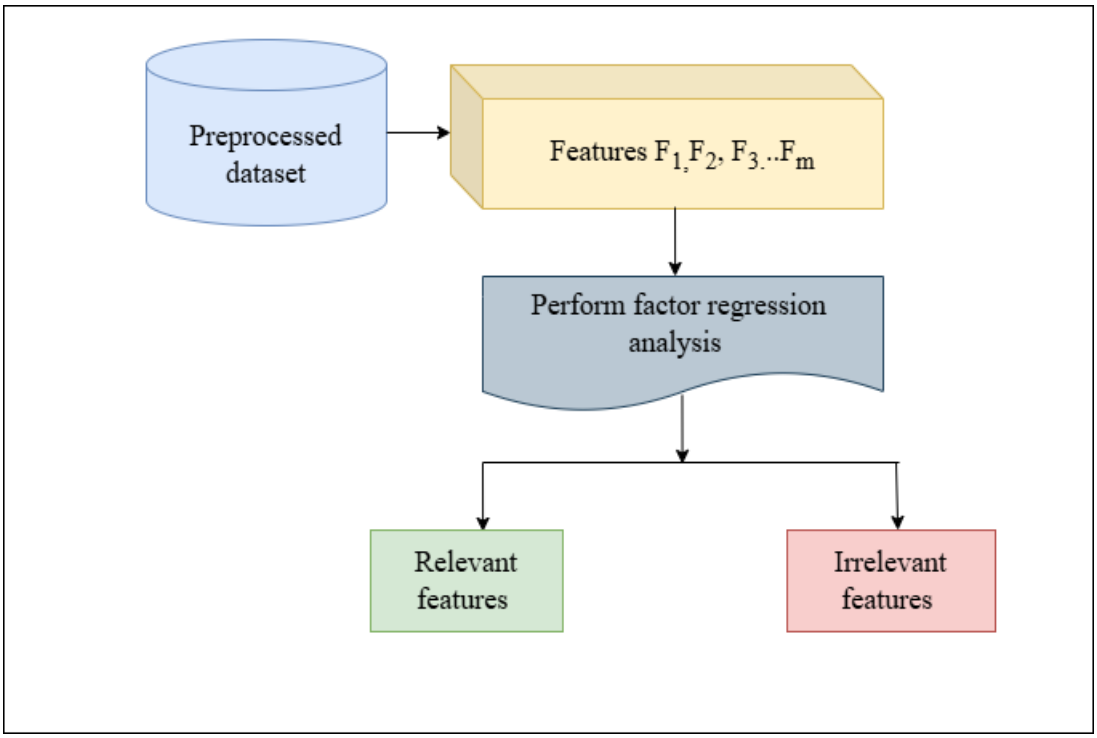



**Figure 2 Flow Process of Feature Selection**

Figure 2 flow process of the feature selection using Tanimoto indexive factor regressive analysis for accurate heart disease

prediction. Let us consider the number of features $F_1, F_2, \dots, F_m$ in the given dataset. Then the factor regression analysis is employed to measure the relationship between the independent variables i.e. features and dependent variable (i.e. target) as follows,

$$R = AL_m + BF_m + C + \epsilon \quad (5)$$

Where, $R$ denotes an output of regression, $DP_1, DP_2, DP_3, \dots DP_n$ denotes a number of data samples or instances, $A, B$ denotes a regression coefficients, $\epsilon$ indicates the error term , $C$ indicates a constant, $L_m$ denotes a latent factor, $F_m$ denotes a observed design factors (input features). The latent factors are measured from the relationships between the features and target. This relationship is measured based on the Tanimoto similarity index as given below,

$$TS = \frac{\sum F_j T_k}{\sqrt{\sum {F_j}^2} + \sqrt{\sum {T_k}^2} - \sum F_j T_k} \quad (6)$$

Where '$TS$' symbolizes the Tanimoto similarity coefficient, $F_j$ denotes an feature, $T_k$ denotes a target variable, $\sum {F_j}^2$ denotes a sum of the squared score of $F_j$, $\sum {T_k}^2$ indicates a sum of the squared score of the $T_k$, $\sum F_j T_k$ denotes a sum of the product of the paired score of $F_j$ and $T_k$. Therefore, Tanimoto similarity coefficient provides the output ranges from 0 to 1.

$$Y = \begin{cases} R > T\ ; & relevant\ features \\ otherwise\ ; & irrelevant\ features \end{cases} \quad (7)$$

If the regression outcomes '$R$' exceeds the threshold '$T$', feature is termed as relevant. Otherwise, the features are irrelevant. Finally, the relevant features are selected for accurate heart disease prediction and removed the other irrelevant features from the dataset. The algorithm for Tanimoto indexive factor regressive analysis is given below,

**Algorithm 2:** Tanimoto indexive factor regressive analysis

**Input**: preprocessed Datasets '$DS$', features $F = \{F_1, F_2, \dots, F_m\}$, patient data samples $PD = \{PD_1, PD_2, \dots, PD_n\}$

**Output**: select relevant features

**Begin**

**1:** Collect the preprocessed dataset as input

**2.** **For** each feature '$F_j$'

3. Measure the regression analysis using (5)

4. Measure the similarity using (6)

5. ***if*** $(Y > T)$ **then**

6. Features are identified as relevant

7. **else**

8. Features are identified as irrelevant

9. **End if**

**10.** Select the relevant features and remove other features

**11.** **end for**

**End**

Algorithm 2 describes the process of relevant feature selection using Tanimoto indexive factor regressive analysis with the aim of improving heart disease prediction while minimizing time consumption. The preprocessed dataset is considered as input for this analysis. Then applying a regression analysis based on Tanimoto similarity coefficient. This similarity measure distinguishes the relevant and irrelevant features with higher accuracy by means of setting the threshold within the dataset. Finally, the relevant features are selected and it listed in table 2 for accurate heart disease prediction in healthcare applications.

**Table 2 Selected Features List**

| S.NO. | Selected relevant features |
|---|---|
| 1 | Age |
| 2 | Gender |

| | |
|---|---|
| 3 | Patient Height |
| 4 | Patient Weight |
| 5 | Cholesterol Levels |
| 6 | Systolic Blood Pressure |
| 7 | Diastolic blood pressure |
| 8 | Glucose level |

### 3.4 Quadratic Discriminant Emphasis boosting ensemble classifier

After the feature selection, classification is performed in SRQE-Boost model for the diagnosis of heart disease using a Quadratic Discriminant **E**mphasis **B**oosting ensemble classifier model with a set of selected features. An **E**mphasis **B**oosting method is an ensemble machine learning algorithm that combines weak classifier to improve strong predictive performance. It works by sequentially training multiple weak classifiers. A weak classifier provides slightly correlated with the true classification, while a strong classifier provides the accurate classification of individuals with and without heart disease. Therefore, the proposed SRQE-Boost model adopts this ensemble approach to effectively distinguish disease presence or absence, thereby increasing prediction accuracy.

In the contrast to other ensemble method, Emphasis boosting algorithms is used for achieving higher accuracy by focusing on misclassified or borderline cases. It also improves the model performance, especially when dealing with complex datasets or imbalanced classes. This Emphasis boosting algorithm produces the final classification output, such as whether a disease is present or absent.

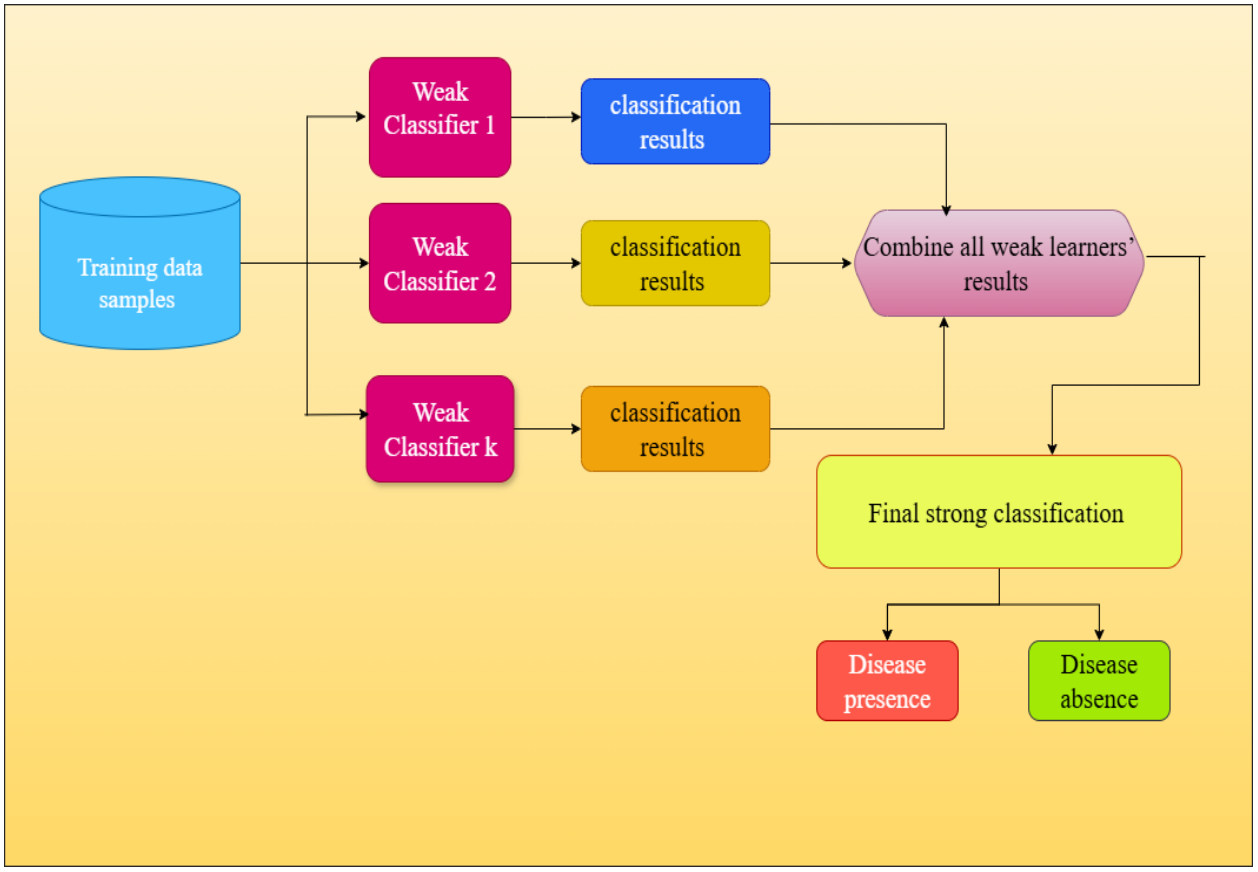



**Figure 3 Schematic Construction of Quadratic Discriminant Emphasis Boost Ensemble Classifier**

Figure 3 illustrates the schematic illustration of Emphasis Boost ensemble classifier for prediction the heart disease presence or absence with higher accuracy and minimum time consumption. The emphasis boosting technique considerers the training set $\{ PD_i, Y_i \}$ where $X_i$ indicates the selected features with training patient samples and $Y_i$ indicates the ensemble

classification output. First, the ensemble boosting technique constructs '$k$' number of weak learners $W_1, W_2, W_3, \ldots . W_k$ as Quadratic Discriminant classifier assumes that each class follows with its own mean and covariance matrix.

Let us consider the number of selected features $F = \{F_1, F_2, \ldots, F_b\}$ with patient data samples $PD = \{PD_1, PD_2, \ldots, PD_n\}$. First, identify the number of classes in the dataset. Then compute the mean vector for each class as follows,

$$\mu_c = \frac{1}{n} \sum_{i=1}^{n} PD_i \quad (8)$$

Where, $\mu_c$ denotes a class mean vector, $n$ denotes a number of data samples in particular class. Then the class covariance vector is computed based on mean value using Gaussian function.

$$CV = \frac{1}{(2\pi\, d)^{-1}} \exp[-0.5 * (PD_i - \mu_c)^T (PD_i - \mu_c)] \quad (9)$$

Where, $CV$ denotes a covariance, $d$ denotes a deviation, $\mu_c$ denotes a mean of particular class, $PD_i$ denotes a patient data samples. With the estimated mean and covariance value, the quadratic discriminant function for particular class is calculated as follows,

$$\varphi_f = -\frac{1}{2} \log |CV_c| - \frac{1}{2} (PD_i - \mu_c)^T CV^{-1} (PD_i - \mu_c) + \log P(c) \quad (10)$$

Where, $\varphi_f$ quadratic discriminant function, $\mu_c$ denotes a **mean vector** for class c, and $CV_c$ denotes a covariance matrix for class c, capturing the spread and shape of the data distribution for that class. The first term, $-\frac{1}{2} \log |CV_c|$ indicates models with large uncertainty or variance. The second term, $-\frac{1}{2} (PD_i - \mu_c)^T \; CV^{-1} (PD_i - \mu_c)$ indicates the Mahalanobis distance between the input $PD_i$and the class mean $\mu_c$, effectively quantifying how far $\mu_c$is from the mean of class. The final term, $\log P(c)$ indicates prior probability of class c. Finally, the classification of data samples is done with the highest discriminant score.

$$Z = \arg\max \varphi_f \quad (11)$$

Where, $Z$ denotes a output of quadratic classifier, $\arg max$ denotes a argument of maximum function, $\varphi_f$ denotes a quadratic discriminant score. It shows that the patient data samples '$PD_i$' is assigned to the particular class with the highest discriminant score. In the heart disease classification, the discriminant function $\varphi_f$ is used to determine whether a given patient data sample'$PD_i$ belongs to the class indicating presence or absence of heart disease. For each class (i.e. heart disease presence = 1, heart disease absence = 0), the classifier computes a discriminant score. The patient sample '$PD_i$ is then assigned to the class with the highest discriminant score. This process allows the model to make accurate and predictions whether a patient is probable to have heart disease or not based on their medical data. In this way, the weak learner classifies the patient's data samples into disease presence or absence. In order to obtain the strong classification output, the weak classification results are combined as follows,

$$Y = \sum_{i=1}^{k} Z_i \quad (12)$$

Where, $Y$ indicates ensemble classification outcomes, $\sum_{i=1}^{k} Z_i$ represents weak classification result. For each output, weights are randomly assigned.

$$Y = \sum_{i=1}^{k} Z_i \; \vartheta_i \quad (13)$$

Where, '$\vartheta_i$represents weights. The proposed boosting technique utilizes the weighted emphasis function to measure the error of classification results obtained from the weak learners,

$$ES = \exp\left[H\left(\left(\sum_{i=1}^{k} Z_i \; \vartheta_i - Y\right)^2 - (1 - H)\left(\sum_{i=1}^{k} Z_i\right)^2\right)\right] \quad (14)$$

Where, $ES$ dentoes a weighted emphasis function, $H$ denotes a weighting constraint ($H = 1$), $Y$ depicts actual classification results, '$\sum_{i=1}^{k} Z_i \; Q_i$' indicates a predicted classification results with weight $Q_i$ and without weight $\sum_{i=1}^{k} Z_i$. From the (10), by substituting '$H$' value is 1 and obtain the final strong classification output,

$$ES = \exp\left[\left(\sum_{i=1}^{k} Z_i \; \vartheta_i - Y\right)^2\right] \quad (15)$$

According to the estimated error value, the weak learner weight gets updated. By applying a damped least-squares method, the classification results are obtained by finding objective function i.e. minimum error value.

$$F = \arg\min\left[\exp\left[\left(\sum_{i=1}^{k} Z_i \; \vartheta_i - Y\right)^2\right]\right] \quad (16)$$

Where, $F$ denotes an output of damped least-squares method, $arg\ min$ denotes an argument of minimum function. Finally, the strong learner results with minimum error are considered as the final strong classified result. Based on the classification results, patients with heart disease presence or absence are correctly detected. The Emphasis Boost classification algorithm is given below,

**// Algorithm 3: Quadratic Discriminant Emphasis Boosting ensemble classifier**

**Input**: Selected relevant features $F = \{F_1, F_2, \dots, F_b\}$, patient data samples $PD = \{PD_1, PD_2, \dots, PD_n\}$

**Output**: Improve the disease prediction accuracy

**Begin**

// **Initialize the classes** $c_1 - disease\ presence$ (1) , $c_2 - disease\ absence$ (0)

**1**: **For each** patient data samples with selected features

2. Construct '$k$' number of weak classifier

3. **For each** class 'c'

4. Compute mean vector '$\mu_c$' using (8)

5. Compute covariance vector '$CV$' using (9)

6. Compute quadratic discriminant function '$\varphi_f$' using (10)

7. ***if*** $\arg\max \varphi_f$ **then**

8. Classify the input samples into disease presence or absence

**9.** **End if**

**10.** **End for**

**11.** **End**

**12.** Combine the set of weak learner results ' $Y = \sum_{i=1}^{k} Z_i$ '

**13.** **for each** weak learner results '$Z_i$'

**14.** Initialize the weight ' $\vartheta_i$'

**15.** Apply the emphasis function using (14)

**16.** Find the weak learner results with minimum error using (16)

**17. end for**

**18.** **Return** (accurate heart disease prediction output)

**End**

Algorithm 3 provided above outlines the step-by-step process of heart disease prediction using Quadratic Discriminant Emphasis Boosting ensemble classifier. This ensemble technique constructs multiple weak learners using the selected relevant features. First, weak learners initialize the classification output. For each patient data samples, mean vector and covariance is computed. Based on the estimated mean and covariance value, Discriminant score is computed. Finally, the maximum Discriminant score is selected for assigning the input sample to the corresponding class. Subsequently, the results from these weak learners are combined, and weight values are initialized. The emphasis function is then applied to measure the error for each weak learner's classification results. Finally, the weak learner with the minimum error is chosen as the final classification outcome. Based on this classification, heart disease prediction is obtained with higher accuracy.

## 4. EXPERIMENTAL SETUP

In this section, experimental assessment of proposed SRQE-Boost model and existing LR [1], stacking-based classification model [2], are implemented in python high level programming language using cardiovascular disease dataset taken from https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset. The main aim is to find the presence and absence of cardiovascular disease (i.e. heart-related disease) in diabetes patients. The dataset consists of 13 attributes and 70000 instances. A cardiovascular disease is a group of disorders of the heart and blood vessels. This dataset is used to help the healthcare professionals in predicting and preventing the risks of heart disease. There are 13 attributes or features are related with each patient, which include a various demographic, medical, and lifestyle-related factors for identifying a cardiovascular i.e. heart disease. The attribute description is listed in table 1.

To perform the experimental evaluation, a random sampling method is applied to select patient data samples from the dataset. This method ensures that the data samples are chosen randomly, allowing for an unbiased calculation of the

performance of both the proposed and existing methods. Random sampling is also known as probability sampling method that involves selecting a random subset of data samples from the population. Based on this approach, the number of patient data samples ranged from 7,000 to 70,000 across ten different runs. For each run, various performance results are observed corresponding to the randomly selected patient data samples.

### 4.1 Implementation scenario

The SRQE-Boost model is experimentally analyzed to measure its performance in heart disease prediction. The evaluation process involves key steps such as data collection, data pre-processing, feature selection and classification. This assessment is conducted using the cardiovascular disease dataset. Initially, patient data sample are collected from the dataset is illustrated in Figure 4.

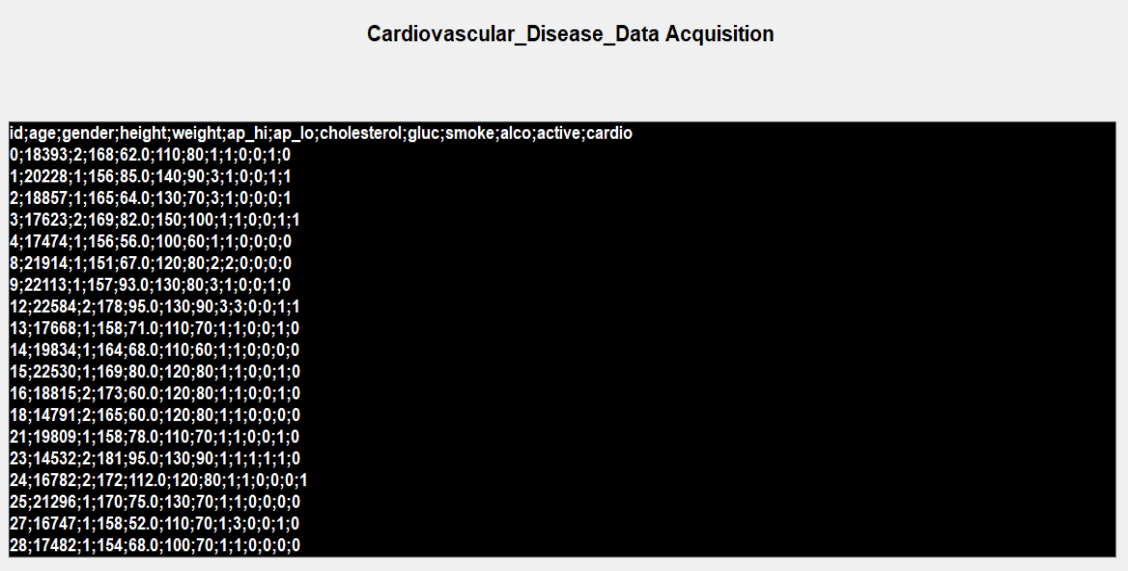
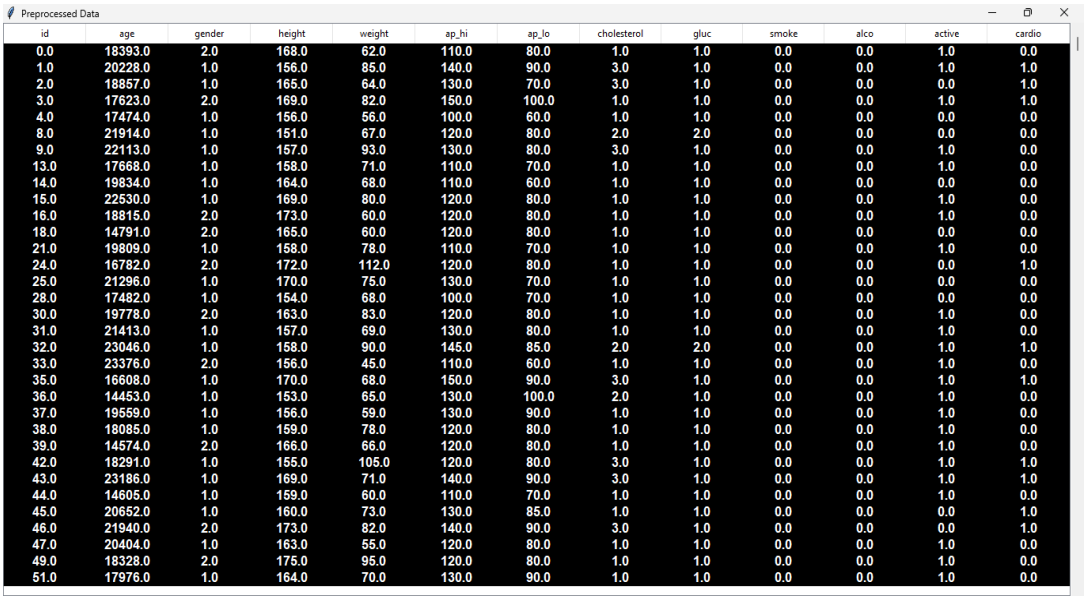



**Figure 4 Sample Data Collection from Dataset**

After data collection, data processing is carried out to handle missing data and outlier data removal for analysis. The original dataset had a size of 2,873 KB. After completing the preprocessing steps, the dataset size was reduced to 2,338kb due to the outlier data removal. The outcomes of the preprocessing phase are illustrated in Figure 5.

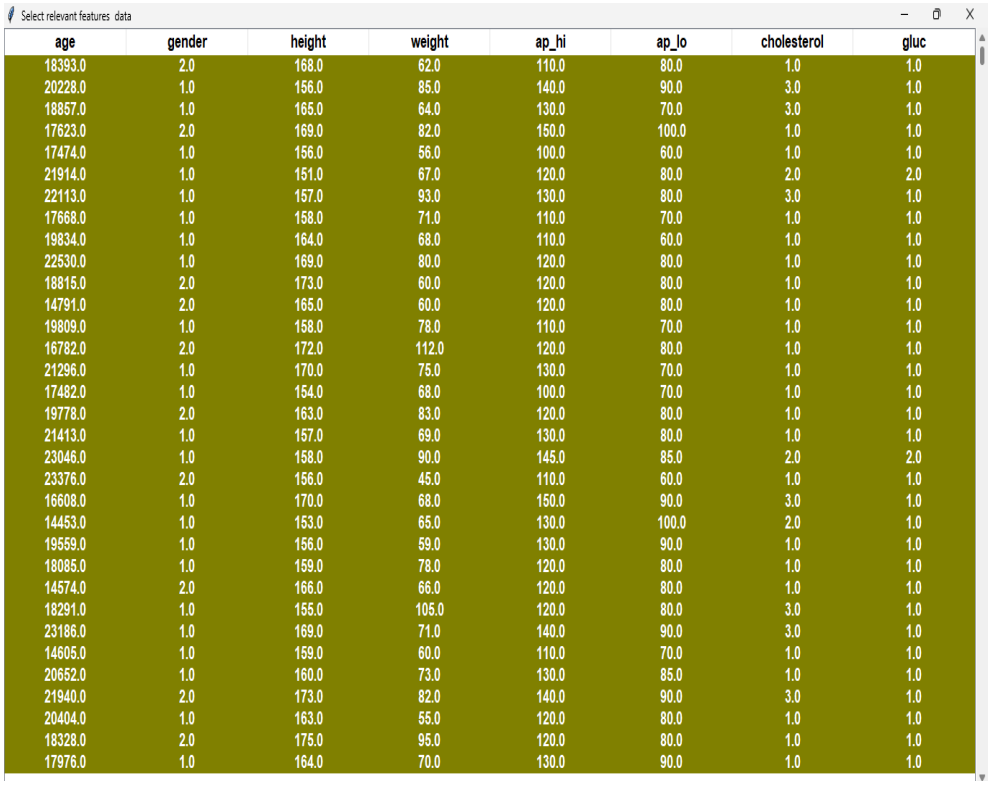



| id | age | gender | height | weight | ap_hi | ap_lo | cholesterol | gluc | smoke | alco | active | cardio |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 18393.0 | 2.0 | 168.0 | 62.0 | 110.0 | 80.0 | 1.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| 1.0 | 20228.0 | 1.0 | 156.0 | 85.0 | 140.0 | 90.0 | 3.0 | 1.0 | 0.0 | 0.0 | 1.0 | 1.0 |
| 2.0 | 18857.0 | 1.0 | 165.0 | 64.0 | 130.0 | 70.0 | 3.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| 3.0 | 17623.0 | 2.0 | 169.0 | 82.0 | 150.0 | 100.0 | 1.0 | 1.0 | 0.0 | 0.0 | 1.0 | 1.0 |
| 4.0 | 17474.0 | 1.0 | 156.0 | 56.0 | 100.0 | 60.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 8.0 | 21914.0 | 1.0 | 151.0 | 67.0 | 120.0 | 80.0 | 2.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 9.0 | 22113.0 | 1.0 | 157.0 | 93.0 | 130.0 | 80.0 | 3.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| 13.0 | 17668.0 | 1.0 | 158.0 | 71.0 | 110.0 | 70.0 | 1.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| 14.0 | 19834.0 | 1.0 | 164.0 | 68.0 | 110.0 | 60.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 15.0 | 22530.0 | 1.0 | 169.0 | 80.0 | 120.0 | 80.0 | 1.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| 16.0 | 18815.0 | 2.0 | 173.0 | 60.0 | 120.0 | 80.0 | 1.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| 18.0 | 14791.0 | 2.0 | 165.0 | 60.0 | 120.0 | 80.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 21.0 | 19809.0 | 1.0 | 158.0 | 78.0 | 110.0 | 70.0 | 1.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| 24.0 | 16782.0 | 2.0 | 172.0 | 112.0 | 120.0 | 80.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| 25.0 | 21296.0 | 1.0 | 170.0 | 75.0 | 130.0 | 70.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 28.0 | 17482.0 | 1.0 | 154.0 | 68.0 | 100.0 | 70.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 30.0 | 19778.0 | 2.0 | 163.0 | 83.0 | 120.0 | 80.0 | 1.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| 31.0 | 21413.0 | 1.0 | 157.0 | 69.0 | 130.0 | 80.0 | 1.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| 32.0 | 23046.0 | 1.0 | 158.0 | 90.0 | 145.0 | 85.0 | 2.0 | 2.0 | 0.0 | 0.0 | 1.0 | 1.0 |
| 33.0 | 23376.0 | 2.0 | 156.0 | 45.0 | 110.0 | 60.0 | 1.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| 35.0 | 16608.0 | 1.0 | 170.0 | 68.0 | 150.0 | 90.0 | 3.0 | 1.0 | 0.0 | 0.0 | 1.0 | 1.0 |
| 36.0 | 14453.0 | 1.0 | 153.0 | 65.0 | 130.0 | 100.0 | 2.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| 37.0 | 19559.0 | 1.0 | 156.0 | 59.0 | 130.0 | 90.0 | 1.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| 38.0 | 18085.0 | 1.0 | 159.0 | 78.0 | 120.0 | 80.0 | 1.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| 39.0 | 14574.0 | 2.0 | 166.0 | 66.0 | 120.0 | 80.0 | 1.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| 42.0 | 18291.0 | 1.0 | 155.0 | 105.0 | 120.0 | 80.0 | 3.0 | 1.0 | 0.0 | 0.0 | 1.0 | 1.0 |
| 43.0 | 23186.0 | 1.0 | 169.0 | 71.0 | 140.0 | 90.0 | 3.0 | 1.0 | 0.0 | 0.0 | 1.0 | 1.0 |
| 44.0 | 14605.0 | 1.0 | 159.0 | 60.0 | 110.0 | 70.0 | 1.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| 45.0 | 20652.0 | 1.0 | 160.0 | 73.0 | 130.0 | 85.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| 46.0 | 21940.0 | 2.0 | 173.0 | 82.0 | 140.0 | 90.0 | 3.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| 47.0 | 20404.0 | 1.0 | 163.0 | 55.0 | 120.0 | 80.0 | 1.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| 49.0 | 18328.0 | 2.0 | 175.0 | 95.0 | 120.0 | 80.0 | 1.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| 51.0 | 17976.0 | 1.0 | 164.0 | 70.0 | 130.0 | 90.0 | 1.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 |

**Figure 5 Preprocessed Dataset**

After preprocessing, identifying the most relevant and informative features from a dataset while eliminating redundant or

irrelevant ones. The goal is to enhance model accuracy, and lower computational time by minimizing the feature space. In this process, SRQE-Boost model selects optimal eight features as shown in figure 10.


Select relevant features data

| age | gender | height | weight | ap_hi | ap_lo | cholesterol | gluc |
|---|---|---|---|---|---|---|---|
| 18393.0 | 2.0 | 168.0 | 62.0 | 110.0 | 80.0 | 1.0 | 1.0 |
| 20228.0 | 1.0 | 156.0 | 85.0 | 140.0 | 90.0 | 3.0 | 1.0 |
| 18857.0 | 1.0 | 165.0 | 64.0 | 130.0 | 70.0 | 3.0 | 1.0 |
| 17623.0 | 2.0 | 169.0 | 82.0 | 150.0 | 100.0 | 1.0 | 1.0 |
| 17474.0 | 1.0 | 156.0 | 56.0 | 100.0 | 60.0 | 1.0 | 1.0 |
| 21914.0 | 1.0 | 151.0 | 67.0 | 120.0 | 80.0 | 2.0 | 2.0 |
| 22113.0 | 1.0 | 157.0 | 93.0 | 130.0 | 80.0 | 3.0 | 1.0 |
| 17668.0 | 1.0 | 158.0 | 71.0 | 110.0 | 70.0 | 1.0 | 1.0 |
| 19834.0 | 1.0 | 164.0 | 68.0 | 110.0 | 60.0 | 1.0 | 1.0 |
| 22530.0 | 1.0 | 169.0 | 80.0 | 120.0 | 80.0 | 1.0 | 1.0 |
| 18815.0 | 2.0 | 173.0 | 60.0 | 120.0 | 80.0 | 1.0 | 1.0 |
| 14791.0 | 2.0 | 165.0 | 60.0 | 120.0 | 80.0 | 1.0 | 1.0 |
| 19809.0 | 1.0 | 158.0 | 78.0 | 110.0 | 70.0 | 1.0 | 1.0 |
| 16782.0 | 2.0 | 172.0 | 112.0 | 120.0 | 80.0 | 1.0 | 1.0 |
| 21296.0 | 1.0 | 170.0 | 75.0 | 130.0 | 70.0 | 1.0 | 1.0 |
| 17482.0 | 1.0 | 154.0 | 68.0 | 100.0 | 70.0 | 1.0 | 1.0 |
| 19778.0 | 2.0 | 163.0 | 83.0 | 120.0 | 80.0 | 1.0 | 1.0 |
| 21413.0 | 1.0 | 157.0 | 69.0 | 130.0 | 80.0 | 1.0 | 1.0 |
| 23046.0 | 1.0 | 158.0 | 90.0 | 145.0 | 85.0 | 2.0 | 2.0 |
| 23376.0 | 2.0 | 156.0 | 45.0 | 110.0 | 60.0 | 1.0 | 1.0 |
| 16608.0 | 1.0 | 170.0 | 68.0 | 150.0 | 90.0 | 3.0 | 1.0 |
| 14453.0 | 1.0 | 153.0 | 65.0 | 130.0 | 100.0 | 2.0 | 1.0 |
| 19559.0 | 1.0 | 156.0 | 59.0 | 130.0 | 90.0 | 1.0 | 1.0 |
| 18085.0 | 1.0 | 159.0 | 78.0 | 120.0 | 80.0 | 1.0 | 1.0 |
| 14574.0 | 2.0 | 166.0 | 66.0 | 120.0 | 80.0 | 1.0 | 1.0 |
| 18291.0 | 1.0 | 155.0 | 105.0 | 120.0 | 80.0 | 3.0 | 1.0 |
| 23186.0 | 1.0 | 169.0 | 71.0 | 140.0 | 90.0 | 3.0 | 1.0 |
| 14605.0 | 1.0 | 159.0 | 60.0 | 110.0 | 70.0 | 1.0 | 1.0 |
| 20652.0 | 1.0 | 160.0 | 73.0 | 130.0 | 85.0 | 1.0 | 1.0 |
| 21940.0 | 2.0 | 173.0 | 82.0 | 140.0 | 90.0 | 3.0 | 1.0 |
| 20404.0 | 1.0 | 163.0 | 55.0 | 120.0 | 80.0 | 1.0 | 1.0 |
| 18328.0 | 2.0 | 175.0 | 95.0 | 120.0 | 80.0 | 1.0 | 1.0 |
| 17976.0 | 1.0 | 164.0 | 70.0 | 130.0 | 90.0 | 1.0 | 1.0 |

**Figure 6 Feature Selection Outcomes**

The performance of heart disease prediction is enhanced by identifying the eight most relevant features and eliminating unnecessary or redundant features. By selecting these eight important features, such as age, gender, patient height, weight, cholesterol levels, systolic blood pressure, Diastolic blood pressure and glucose, the SRQE-Boost model becomes more efficient, faster to train, and often more accurate heart disease prediction. These key features are directly correlated to patient heart conditions and determining the disease risk level. Age is a critical factor for developing heart disease risk due to the gradual increase of sign in the arteries, higher blood pressure, and other age-related factors. Gender plays a important role in the risk for heart disease. Men generally have a higher risk of heart disease than the women. The Taller individuals have different body compositions that affect cardiovascular health,

Patient Weight is used in calculating BM related to the risk of developing heart disease. Overweight's are at a higher risk due to high cholesterol levels, high blood pressure, and insulin.

Cholesterol is the most important features of heart disease risk. High levels of cholesterol cause a major cause of heart attack. Systolic blood pressure is the pressure in the arteries. High systolic pressure (hypertension) increases the workload on the heart and damages the arteries, leading to a higher risk of heart disease. The diastolic blood pressure measures the pressure in the arteries between heartbeats. High value of diastolic pressure and glucose levels significantly increase the risk of heart disease.

Finally, **Classification task involves classifying the category or class label such as disease presence or absence with optimally selected features.**

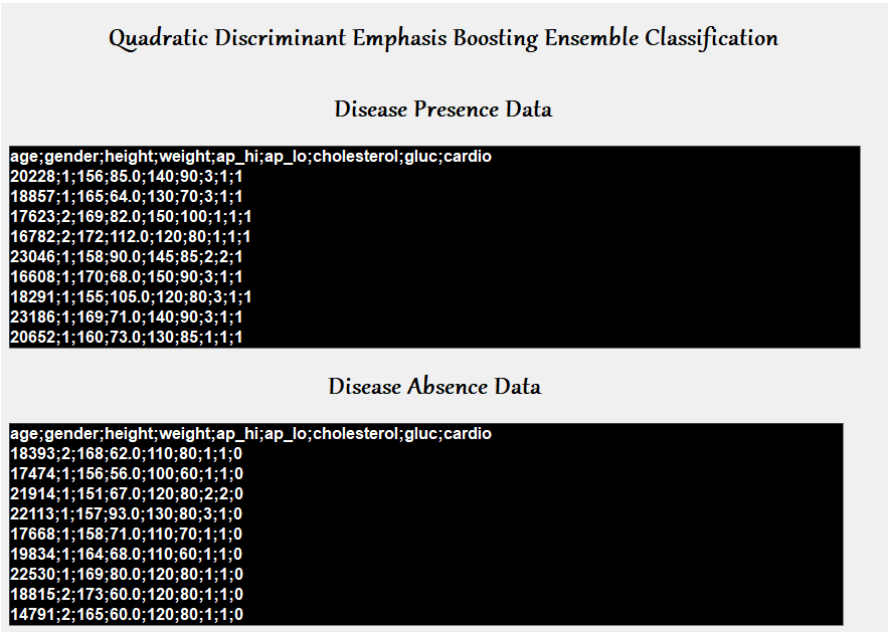



**Figure 7** Classification Outcomes of **SRQE-Boost model**

## 5. PERFORMANCE COMPARISION ANALYSIS

In this section, performance of the proposed SRQE-Boost model and existing LR [1], stacking-based classification model [2] are discussed with various metrics, including feature selection accuracy, precision, recall, F1 score, specificity, AUC, MCC, prediction time and memory consumption, confusion matrix with different number of data samples.

**Prediction accuracy**: Prediction accuracy refers to the model's ability to correctly detect the presence or absence of heart disease. It is a critical performance metric that evaluates the effectiveness of the classification process. The overall accuracy of the model is mathematically computed using the following formula:

$$\boldsymbol{PA} = \left(\frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}}\right) * \mathbf{100} \qquad (17)$$

Where, $\boldsymbol{PA}$ denotes an prediction accuracy, true positive '$\boldsymbol{TP}$' indicates correctly predicted presence of heart disease, true negative '$\boldsymbol{TN}$' denotes the correctly predicted absence of heart disease, false positive '$\boldsymbol{FP}$' represents the incorrectly predicted presence of heart disease, false negative '$\boldsymbol{FN}$' represents incorrectly predicted absence of heart disease. It is measured in percentage (%).

Precision: it refers to the model's ability to accurately detect the presence of heart disease among all patient samples. The precision is mathematically computed as follows,

$$\boldsymbol{Precision} = \left(\frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}}\right) \qquad (18)$$

Where, $TP$ denotes the true positive, $FP$ represents the false positive.

**Recall:** it also known as Sensitivity, is a performance metric used to measure a model's ability to correctly detect the presence of heart disease. The recall is mathematically calculated as follows,

$$\boldsymbol{Recall} = \left(\frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}}\right) \qquad (19)$$

Where, $TP$ denotes the true positive, $FN$ represents the false negative.

**F1 score:** it is a common evaluation metric used in heart disease prediction tasks. It is a measure of harmonic mean of the precision and recall.

$$\boldsymbol{F1\ score} = \left(\mathbf{2} * \frac{\boldsymbol{precision*recall}}{\boldsymbol{precision*recall}}\right) \qquad (20)$$

**Specificity:** it plays a crucial role in heart disease prediction, particularly in reducing false positives. It measures the model's ability to correctly identify healthy patients. The formula for computing the specificity is expressed as follows,

$$Specificity = \left(\frac{\boldsymbol{TN}}{\boldsymbol{TN+FP}}\right) \quad \mathbf{(21)}$$

Where, $TN$ denotes the true negative, $FP$ represents the false positive.

**Mathew Correlation Coefficient (MCC):** it is a robust statistical analysis that provides a evaluation of a machine learning model performance by considering all four categories of the confusion matrix such as true positives, false negatives, true negatives, and false positives. It provides the output value from 0 to 1. The formula for computing the MMC is mathematically expressed as follow.

$$MCC = \left(\frac{(TP*TN)-(FP*FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}\right) \quad (22)$$

**Prediction time**: It is measured as an amount of time taken by the method for predicting the heart disease with respect to number of input data samples. The overall time consumption is measured as follows,

$$\boldsymbol{PT} = \sum_{i=1}^{n} \boldsymbol{PD_i} * \boldsymbol{TM(DP)} \quad (23)$$

Where, $\boldsymbol{PT}$ denotes a prediction time based on the patient data samples '$PD_i$' and the actual time consumed in predicting the heart disease denoted by '$\boldsymbol{TM(DP)}$'. It is measured in terms of milliseconds (ms).

Memory consumption: It refers to as an amount of memory space consumed by algorithm for heart disease prediction. The overall memory consumption is computed as follows,

$$\boldsymbol{MC} = \sum_{i=1}^{n} \boldsymbol{PD_i} * \boldsymbol{Mem(DP)} \quad (24)$$

Where, $MC$ denotes a memory consumption based on the patient data '$PD_i$' and the memory space consumed in heart disease prediction denoted by '$\boldsymbol{Mem(DP)}$'. It is measured in terms of Kilobytes (KB).

**Table 3 Comparison of Prediction Accuracy**

| Number of patient data | Prediction accuracy (%) (without feature selection) | | | Prediction accuracy (%) (with feature selection) | | |
|---|---|---|---|---|---|---|
| | Proposed SRQE-Boost | LR [1] | stacking-based classification model [2] | Proposed SRQE-Boost | LR [1] | stacking-based classification model [2] |
| **7000** | 94.22 | 88.57 | 90.71 | 95.60 | 89.57 | 91.14 |
| **14000** | 94.74 | 88.65 | 90.55 | 95.36 | 89.36 | 91.08 |
| **21000** | 93.65 | 88.66 | 90.41 | 95.85 | 89.45 | 91.12 |
| **28000** | 93.94 | 88.25 | 90.26 | 95.74 | 89.47 | 91.45 |
| **35000** | 94.75 | 88.57 | 90.32 | 95.36 | 89.63 | 91.63 |
| **42000** | 94.82 | 88.74 | 90.14 | 95.33 | 89.44 | 91.74 |
| **49000** | 93.89 | 87.89 | 90.36 | 95.78 | 89.63 | 91.36 |
| **56000** | 93.89 | 88.12 | 90.74 | 95.36 | 89.32 | 91.74 |
| **63000** | 94.02 | 88.74 | 90.65 | 95.63 | 89.77 | 91.33 |
| **70000** | 94.08 | 88.25 | 90.65 | 95.33 | 89.05 | 91.21 |

Table 3 describes the experimental results of disease prediction accuracy with and without feature selection along with the number of patient data taken in the ranges from 7000 to 70000 taken from the dataset. The disease prediction accuracy is measured using three methods namely SRQE-Boost model and existing LR [1], stacking-based classification model [2]. The observed performance results show that the accuracy of SRQE-Boost model was higher than the different feature selection schemes. For each method, ten different performance results were observed and compared. The overall comparative analysis shows that the disease prediction accuracy using the SRQE-Boost model with feature selection increased significantly by 7% and 5% compared to methods proposed in [1] and [2], respectively. Similarly, the SRQE-Boost model without feature selection also demonstrated improved prediction accuracy, with an increase of 6% and 4% compared to [1] and [2], respectively.

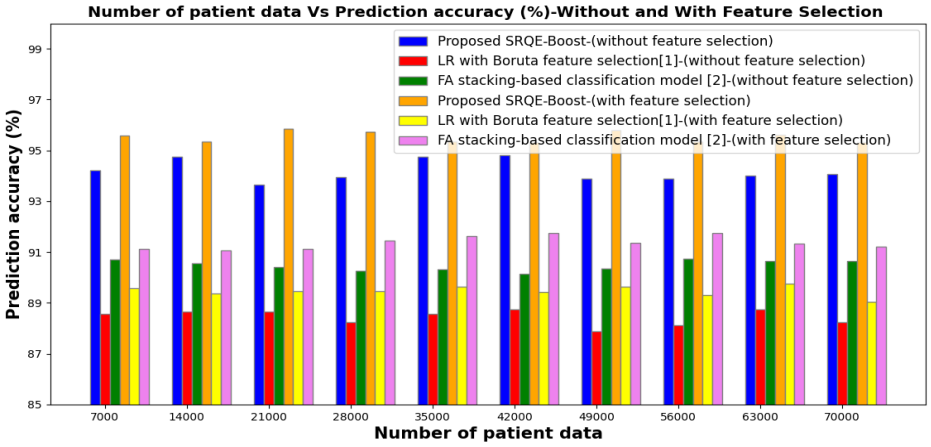



**Figure 8 Performance Analysis of Prediction Accuracy**

Figure 8 illustrates performance analysis of feature selection accuracy using three methods namely SRQE-Boost model and existing LR [1], stacking-based classification model [2]. As shown in figure 8, the horizontal axis illustrates the number of patient data ranging from 7000 to 70000, while the vertical axis indicates the performance outcomes of feature selection accuracy. Among three methods, the performance of SRQE-Boost model is better when compared to the existing feature section methods. This is because of applying the Quadratic Discriminant Emphasis Boosting ensemble classifier. This ensemble technique constructs Quadratic Discriminant weak classifier for classifying the patient data into disease presence or absence. Subsequently, the results from these weak learners are combined and find the best weak learner results with minimum error. This capability of SRQE-Boost model increases the performance of disease prediction accuracy.

**Table 4 Comparison of Precision**

| **Number of patient data** | **Precision (without feature selection)** | | | **Precision (with feature selection)** | | |
|---|---|---|---|---|---|---|
| | Proposed SRQE-Boost | LR [1] | stacking-based classification model [2] | Proposed SRQE-Boost | LR [1] | stacking-based classification model [2] |
| **7000** | 0.928 | 0.906 | 0.915 | 0.935 | 0.918 | 0.931 |
| **14000** | 0.921 | 0.904 | 0.911 | 0.934 | 0.916 | 0.921 |
| **21000** | 0.927 | 0.905 | 0.909 | 0.933 | 0.914 | 0.922 |
| **28000** | 0.922 | 0.906 | 0.912 | 0.938 | 0.913 | 0.92 |
| **35000** | 0.922 | 0.904 | 0.908 | 0.939 | 0.912 | 0.923 |
| **42000** | 0.928 | 0.901 | 0.909 | 0.936 | 0.911 | 0.921 |
| **49000** | 0.924 | 0.9 | 0.911 | 0.937 | 0.914 | 0.924 |
| **56000** | 0.926 | 0.903 | 0.908 | 0.938 | 0.915 | 0.919 |
| **63000** | 0.922 | 0.905 | 0.91 | 0.939 | 0.91 | 0.92 |
| **70000** | 0.924 | 0.903 | 0.908 | 0.937 | 0.911 | 0.919 |

The comparison of precision without and without feature selection of three different methods namely SRQE-Boost model and existing LR [1], FA stacking-based classification model [2] is illustrated in table 4. For the better comparison, the various counts of patient data are taken as input in the ranges from 7000, 14000 … 70000. For each classification method, different performance results were observed with respect to number of patient data. The overall results of the SRQE-Boost

model are compared to the existing methods. The overall comparison results demonstrate that the SRQE-Boost model with feature selection improves precision performance by 3% and 2% compared to [1] and [2], respectively. Similarly, the results indicate that the SRQE-Boost model without feature selection increases precision performance by 2% and 1% compared to [1] and [2], respectively. The graphical illustration of the precision of three methods is shown in figure 9.

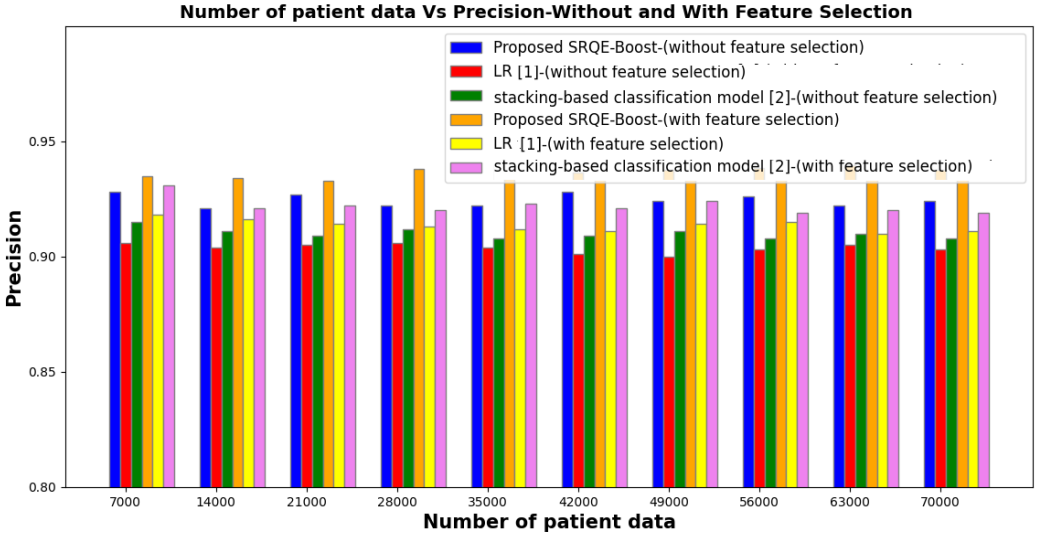



**Figure 9 Performance Analysis of Precision**

Figure 9 shows the performance analysis of precision with and without feature selection related to heart disease with respect to number of patient data samples ranged from 7000 to 70000. Three methods were employed to estimate the precision using SRQE-Boost model and existing LR [1], stacking-based classification model [2]. As shown in figure 9, the number of patient data is considered in horizontal axis, while performance of precision results observed on vertical axis. The evaluation results ensure that the SRQE-Boost model achieved improved precision results compared to the other classification methods. These improved performances of SRQE-Boost model were achieved due to the application of the damped least-squares method in emphases boosting ensemble classification method. The method finds the classification outcomes with minimal error thereby increasing the true positive and minimizing the false positive results in detecting the relevant and irrelevant features.

**Table 5 Comparison of Recall**

| **Number of patient data** | **Recall (without feature selection)** | | | **Recall (with feature selection)** | | |
|---|---|---|---|---|---|---|
| | Proposed SRQE-Boost | LR [1] | stacking-based classification model [2] | Proposed SRQE-Boost | LR [1] | stacking-based classification model [2] |
| **7000** | 0.954 | 0.915 | 0.929 | 0.960 | 0.924 | 0.935 |
| **14000** | 0.952 | 0.914 | 0.928 | 0.962 | 0.922 | 0.934 |
| **21000** | 0.953 | 0.913 | 0.927 | 0.963 | 0.919 | 0.935 |
| **28000** | 0.948 | 0.911 | 0.929 | 0.959 | 0.922 | 0.936 |
| **35000** | 0.953 | 0.918 | 0.931 | 0.961 | 0.925 | 0.937 |
| **42000** | 0.955 | 0.919 | 0.929 | 0.96 | 0.922 | 0.935 |
| **49000** | 0.949 | 0.917 | 0.928 | 0.962 | 0.924 | 0.936 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **56000** | 0.946 | 0.919 | 0.931 | 0.962 | 0.927 | 0.933 |
| **63000** | 0.951 | 0.92 | 0.928 | 0.961 | 0.926 | 0.934 |
| **70000** | 0.95 | 0.922 | 0.929 | 0.959 | 0.925 | 0.933 |

Table 5 illustrates the comparisons results of the recall with and without feature selection versus number of patient data. The above comparison results demonstrate that the performance of recall is said to be higher using proposed SRQE-Boost model and existing LR [1], stacking-based classification model [2]. To enable a more comprehensive comparison, varying sizes of patient data were considered, ranging from 7,000 to 70,000. The comparison of three different methods proves that the performance of recall with feature selection using SRQE-Boost model is considerably increased by 4% than the [1] and also improved by 3% when compared to [2] respectively. Likewise, the SRQE-Boost model without feature selection improves precision performance by 4% and 2% compared to [1] and [2], respectively. The graphical results of recall are shown in the figure 10.

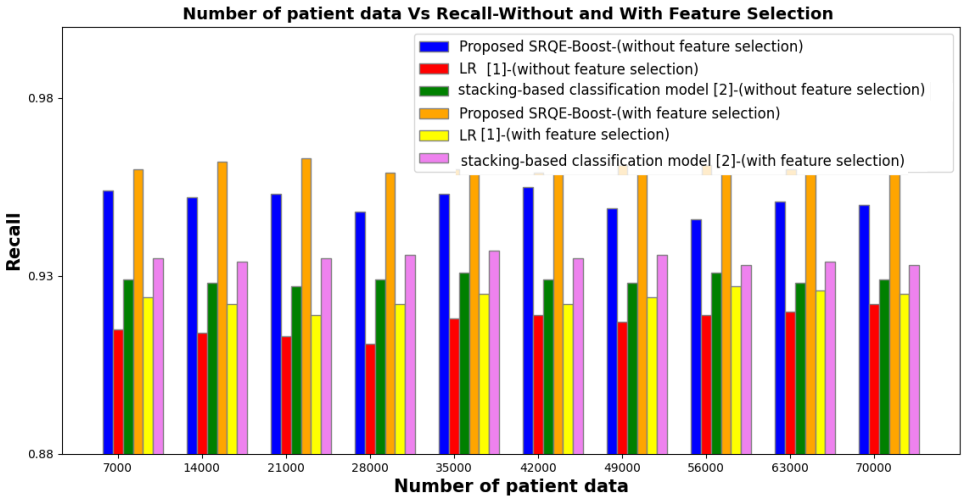



**Figure 10 Performance Analysis of Recall**

Figure 10 given above illustrates the performance outcomes of recall with and without feature selection versus the number of patient data, ranging from 7,000 to 70,000, for three methods namely SRQE-Boost model and existing LR [1], stacking-based classification model [2]. These methods were employed to evaluate performance of recall in feature selection. The horizontal axis denotes the number of patient data, while the vertical axis represents recall performance. Among the three methods, proposed SRQE-Boost model exhibits comparatively better recall performance than [1], [2], respectively. This is because of the SRQE-Boost model utilizes the ensemble classification method for distinguishing the patient data into disease presence or absence while improving the true positive and minimizing false negative.

**Table 6 Comparison of F1 Score**

| **Number of patient data** | **F1 score (without feature selection)** | | | **F1 score (with feature selection)** | | |
|---|---|---|---|---|---|---|
| | Proposed SRQE-Boost | LR [1] | stacking-based classification model [2] | Proposed SRQE-Boost | LR [1] | stacking-based classification model [2] |
| **7000** | 0.940 | 0.910 | 0.921 | 0.947 | 0.920 | 0.932 |
| **14000** | 0.936 | 0.908 | 0.919 | 0.947 | 0.918 | 0.927 |
| **21000** | 0.939 | 0.908 | 0.917 | 0.947 | 0.916 | 0.928 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **28000** | 0.934 | 0.908 | 0.920 | 0.948 | 0.917 | 0.927 |
| **35000** | 0.937 | 0.910 | 0.919 | 0.949 | 0.918 | 0.929 |
| **42000** | 0.941 | 0.909 | 0.918 | 0.947 | 0.916 | 0.927 |
| **49000** | 0.936 | 0.908 | 0.919 | 0.949 | 0.918 | 0.929 |
| **56000** | 0.935 | 0.910 | 0.919 | 0.949 | 0.920 | 0.925 |
| **63000** | 0.936 | 0.912 | 0.918 | 0.949 | 0.917 | 0.926 |
| **70000** | 0.936 | 0.912 | 0.918 | 0.947 | 0.917 | 0.925 |

The performance comparison of F1 scores with and without feature selection is measured using three different methods namely SRQE-Boost model and existing [1], [2] are illustrated in table 6. For the better comparison, the various counts of patient data are taken as input in the ranges from 7000, 14000, 21000 …70,000. For the different counts of input patient data, three various F1 score results were obtained as shown in table 6. Different performance results were observed with different counts of input samples. The overall observed results of the SRQE-Boost model are compared to the existing methods. The overall comparison results prove that the performance of F1 score with and without feature selection is significantly improved by 3% than the [1] and also improved by 2% when compared to [2] respectively. The graphical results of F1 score are shown in the figure 11.

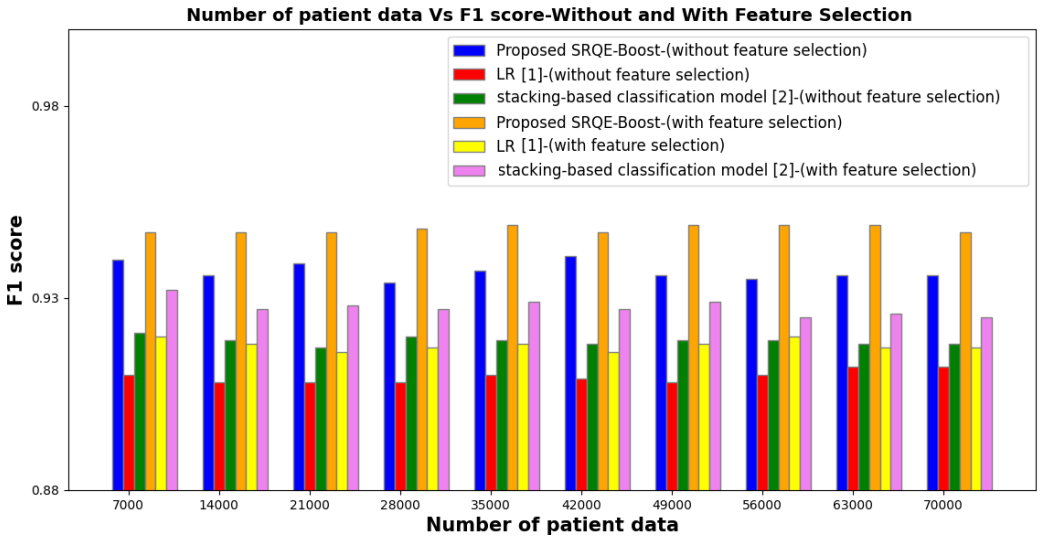



**Figure 11 Performance Analysis of F1 Score**

Figure 11 illustrates the performance results of the F1-score versus number of patient data using three methods namely SRQE-Boost model and existing [1], [2]. The F1-score performance results were measured based on both precision and recall. It is evident from these results that the proposed SRQE-Boost model outperforms the existing methods in terms of F1-score. The application of the SRQE-Boost model enhances both precision and recall during feature selection. This improvement is achieved due to the effective application of the ensemble classification methods, which leads to a higher F1-score.

**Table 7 Comparison of Specificity**

| | **specificity (without feature selection)** | **specificity (with feature selection)** |
|---|---|---|

| **Number of patient data** | Proposed SRQE-Boost | LR [1] | stacking-based classification model [2] | Proposed SRQE-Boost | LR [1] | Stacking-based classification model [2] |
|---|---|---|---|---|---|---|
| **7000** | 0.858 | 0.835 | 0.85 | 0.877 | 0.84 | 0.863 |
| **14000** | 0.857 | 0.833 | 0.845 | 0.875 | 0.838 | 0.862 |
| **21000** | 0.862 | 0.825 | 0.842 | 0.88 | 0.836 | 0.858 |
| **28000** | 0.859 | 0.827 | 0.847 | 0.879 | 0.832 | 0.862 |
| **35000** | 0.858 | 0.826 | 0.846 | 0.876 | 0.832 | 0.862 |
| **42000** | 0.859 | 0.824 | 0.848 | 0.874 | 0.837 | 0.857 |
| **49000** | 0.854 | 0.822 | 0.844 | 0.873 | 0.839 | 0.856 |
| **56000** | 0.859 | 0.823 | 0.843 | 0.877 | 0.832 | 0.857 |
| **63000** | 0.852 | 0.828 | 0.847 | 0.874 | 0.836 | 0.856 |
| **70000** | 0.853 | 0.832 | 0.845 | 0.875 | 0.842 | 0.854 |

Table 7 describes the performance investigation of specificity with and without feature selection against the number of patient data, ranging from 7000 to 70000, taken from the datasets. The specificity is measured using three different methods namely SRQE-Boost model, [1], [2]. The above results indicate that the performance of proposed SRQE-Boost model achieved better performance than the existing methods. The overall comparative analysis indicates that the proposed SRQE-Boost model consistently outperforms the existing approaches. Specifically, the specificity showed a 5% improvement over method [1] and a 2% improvement over method [2], demonstrating the effectiveness of the proposed SRQE-Boost model with feature selection. Moreover, Specificity improved by 4% compared to method [1] and by 1% compared to method [2], highlighting the effectiveness of the proposed SRQE-Boost model without feature selection.

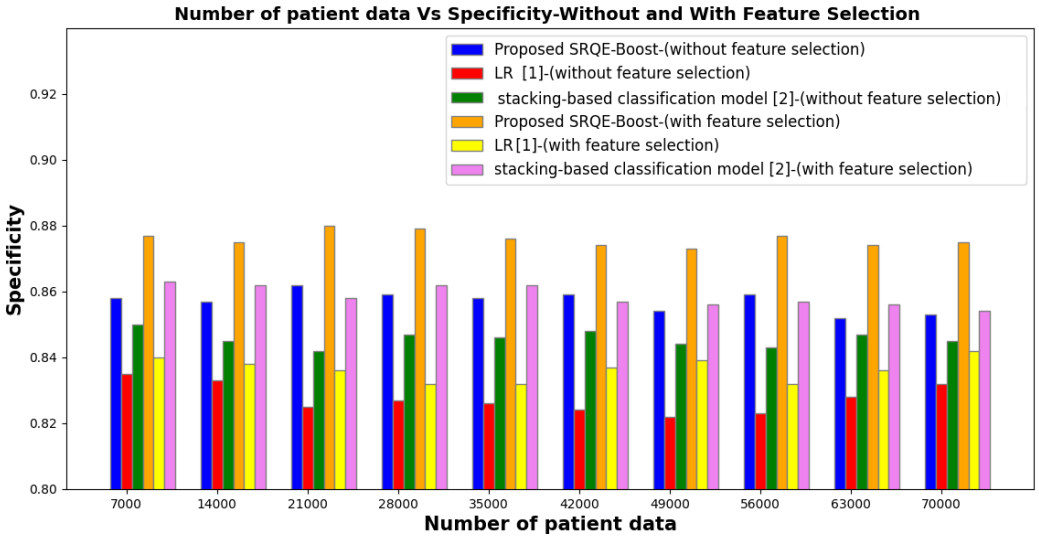



**Figure 12 Performance Analysis of Specificity**

Figures 12 illustrate the performance analysis of specificity with and without feature selection against the number of patient data, ranging from 7000 to 70000, taken from the datasets. The horizontal axis indicates the number of patient data samples, while the vertical axis represents performance outcomes of specificity. The experimental results demonstrate that the SRQE-Boost model achieved higher specificity compared to other two existing SRQE-Boost model, [1], [2]. Various results were observed for each method with different numbers of patient data in accurately determining the relevant

features. The overall results show that the SRQE-Boost model method increases the performance of specificity by reducing the false positives in disease classification process. By detecting the disease classification ensures that the model better distinguishes negative cases, thereby increasing specificity.

**Table 8 Comparison of ROC and AUC Curve**

| False positive rate | True positive rate | | |
|---|---|---|---|
| | Proposed SRQE-Boost | LR [1] | Stacking-based classification model [2] |
| **0** | 0 | 0 | 0 |
| **0.1** | 0.31 | 0.21 | 0.26 |
| **0.2** | 0.55 | 0.32 | 0.38 |
| **0.3** | 0.67 | 0.45 | 0.52 |
| **0.4** | 0.72 | 0.55 | 0.62 |
| **0.5** | 0.84 | 0.63 | 0.72 |
| **0.6** | 0.89 | 0.75 | 0.82 |
| **0.7** | 0.94 | 0.82 | 0.88 |
| **0.8** | 0.96 | 0.86 | 0.9 |
| **0.9** | 0.97 | 0.89 | 0.91 |
| **1** | 0.99 | 0.9 | 0.93 |

Table 8 illustrates the performance comparison of ROC (Receiver Operating Characteristic) and AUC results. In order to evaluate the performance of ROC using proposed SRQE-Boost model, [1], [2] in the relevant and irrelevant feature selection. The ROC results were plotted against True Positive Rate (TPR) versus the False Positive Rate (FPR) at various values settings. The ROC curve for SRQE-Boost model, shows a higher TPR for a given FPR compared to the other classification algorithms, indicating superior performance in disease prediction.

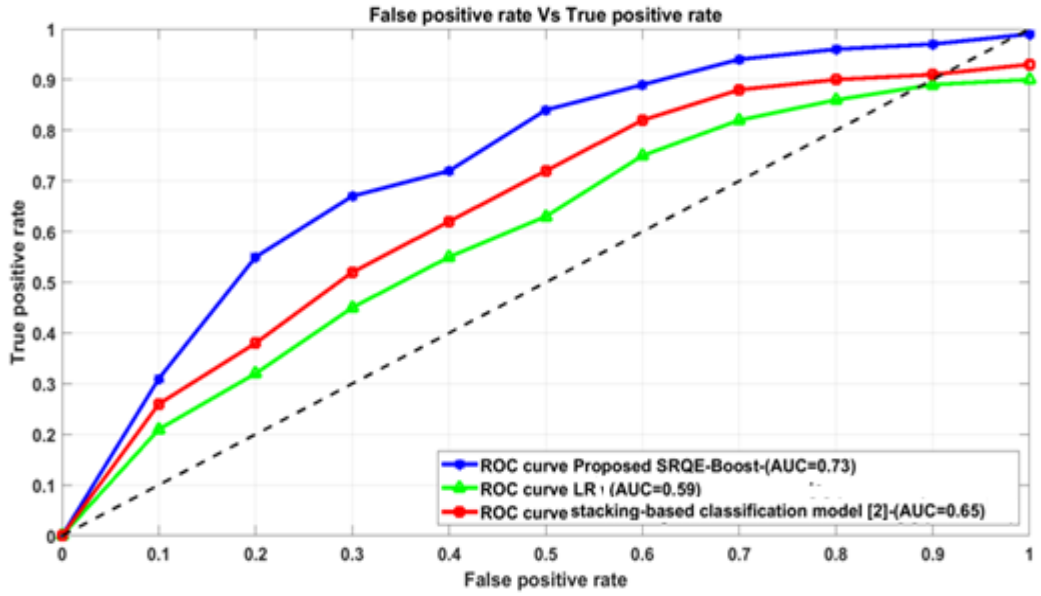



**Figure 13 Performance Analysis of ROC and AUC**

Figure 13 illustrate the ROC results of three methods namely proposed SRQE-Boost model, [1], [2]. It determines the overall ability of the model to accurately distinguish the disease presence and absence. The ROC value ranges from 0 to 1 The Area Under the Curve (AUC) is a performance measurement used to the area under the ROC (Receiver Operating Characteristic) curve. An AUC value is greater than 0.5 indicates a perfect model that accurately distinguishes between disease presence and absence, while an AUC of 0.5 suggests that the model performance is no efficient in disease prediction. AUC values less than 0.5 means that the model performance is poor. In figure 7, the dotted straight line symbolizes a threshold point along the ROC curve, where the model performance is evaluated. Based on the observed results, the final AUC values are 0.73 for the proposed SRQE-Boost model, 0.59 for the existing method [1], and 0.65 for the existing method [2], respectively. These results indicate that all models have very high disease prediction ability. Therefore, the proposed SRQE-Boost model outperforms the existing [1] [2] methods.

**Table 9 Comparison of MCC**

| **Number of patient data** | **MCC** | | |
|---|---|---|---|
| | Proposed SRQE-Boost | LR [1] | Stacking-based classification model [2] |
| **7000** | 0.847 | 0.766 | 0.801 |
| **14000** | 0.905 | 0.785 | 0.865 |
| **21000** | 0.902 | 0.796 | 0.875 |
| **28000** | 0.933 | 0.822 | 0.863 |
| **35000** | 0.95 | 0.833 | 0.886 |
| **42000** | 0.936 | 0.824 | 0.874 |
| **49000** | 0.945 | 0.833 | 0.886 |
| **56000** | 0.911 | 0.826 | 0.862 |
| **63000** | 0.923 | 0.822 | 0.872 |
| **70000** | 0.936 | 0.817 | 0.863 |

Table 9 describes the experimental results of Mathew correlation coefficient by applying three different methods namely SRQE-Boost model, [1], [2]. Among three methods, proposed SRQE-Boost model outperforms well in terms of achieving high MCC results in disease prediction than the other two methods. Let us consider the experimental evaluation containing 7,000 patient data. The proposed SRQE-Boost model demonstrated a notable improvement in MCC, achieving a value of 0.847 during SRQE-Boost model. In contrast, the existing methods [1] and [2] attained MCC scores of 0.766 and 0.801, respectively. The comparative results clearly show that the proposed SRQE-Boost model consistently exceed the existing techniques. Notably, it delivered a 13% increase in MCC compared to method [1] and a 6% gain over method [2], highlight its effectiveness in reducing false positives, false negative and increasing the true positive and true negative.

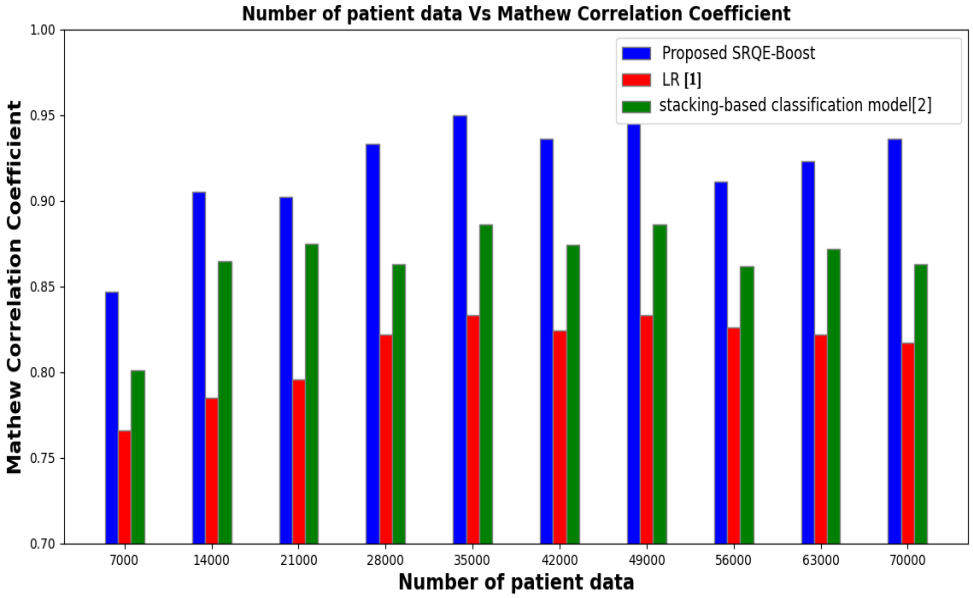



**Figure 14 Performance Analysis of MCC**

Figure 14 illustrates the MCC using three methods namely SRQE-Boost model, [1], [2]. The x-axis represents the number of patient data ranging from 7000 to 70000, while the y-axis shows the MCC observed in disease prediction using three different methods. The results indicate that the SRQE-Boost model enhance the overall performance of MCC compared to [1] and [2]. This higher MCC value help the SRQE-Boost model achieve better balance in predicting both positive and negative classes. The SRQE-Boost model improves MCC by eliminating noisy or irrelevant features, reducing overfitting, and ensuring the high disease prediction accuracy.

**Table 10 Comparison of Prediction Time**

| **Number of patient data** | **prediction time (ms) (without feature selection)** | | | **prediction time (ms) (with feature selection)** | | |
|---|---|---|---|---|---|---|
| | Proposed SRQE-Boost | LR [1] | Stacking-based classification model [2] | Proposed SRQE-Boost | LR [1] | Stacking-based classification model [2] |
| **7000** | 36.8 | 45.6 | 42.5 | 31.8 | 37.5 | 34.8 |
| **14000** | 38.6 | 48.5 | 46.3 | 34.7 | 42.8 | 38.7 |
| **21000** | 43.6 | 50.2 | 48.6 | 36.9 | 45.7 | 40.2 |
| **28000** | 46.8 | 52.8 | 50.2 | 40.2 | 47.6 | 43.6 |
| **35000** | 48.4 | 55.6 | 52.6 | 43.5 | 50.2 | 47.2 |
| **42000** | 51.5 | 60.3 | 55.8 | 47.6 | 53.6 | 50.3 |
| **49000** | 56.8 | 65.7 | 60.3 | 50.2 | 56.4 | 53.8 |
| **56000** | 60.5 | 68.6 | 63.5 | 53.7 | 60.2 | 57.6 |
| **63000** | 63.7 | 72.5 | 67.2 | 55.9 | 63.7 | 60.2 |
| **70000** | 65.2 | 74.6 | 70.5 | 58.3 | 65.8 | 63.7 |

Table 10 describes the experimental outcomes of the prediction time with and without feature section versus a number of patients data collected from the dataset. The feature selection time is measured using three different techniques namely the SRQE-Boost model, [1], [2]. The observed results indicate that the proposed SRQE-Boost model outperforms well than conventional feature selection methods. These experimental results of SRQE-Boost model were then compared to the existing methods. The average value of ten comparison results confirms that the prediction time of the SRQE-Boost model with feature selection is considerably reduced by 14% and 8% when compared to the existing methods [1] and [2] respectively. The average of ten comparison trials confirms that the prediction time of the SRQE-Boost model without feature selection is significantly reduced by 14% and 9% compared to the existing methods [1] and [2] respectively. The graphical analysis of feature selection time is shown in figure 15.

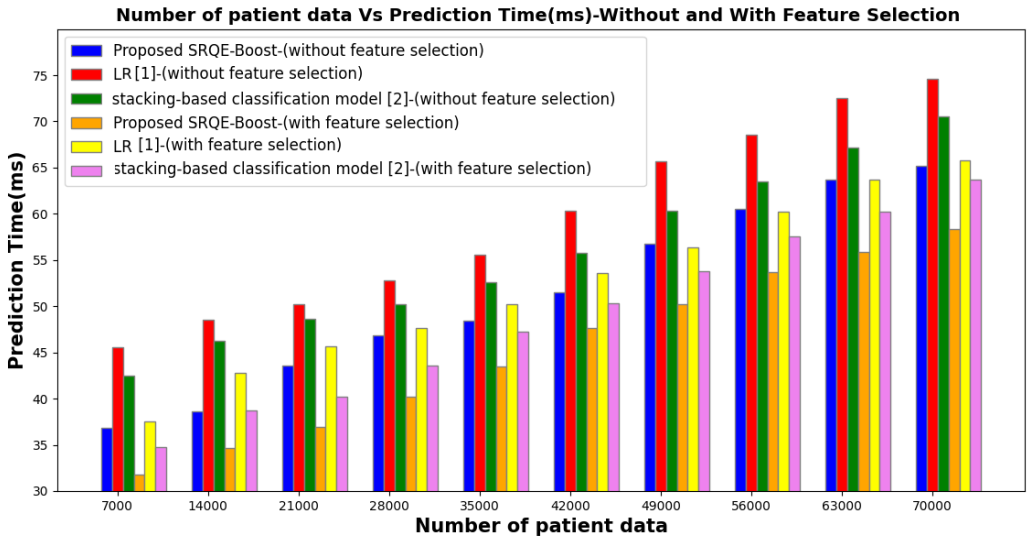



**Figure 15 Performance Analysis of Prediction Time**

Figure 15 illustrates the performance analysis of prediction time with and without feature selection versus the number of patient data ranges from 7000 to 70000. The graphical plot illustrates the prediction time for all three methods progressively increased in a linear manner while increasing the number of patient data. Specifically, the prediction time for the SRQE-Boost model is considerably minimized when compared to the existing methods [1] and [2]. This is due to SRQE-Boost model method performs the data preprocessing and feature selection to handle the missing data using linear spline interpolation method. Furthermore, outlier data is identified through the Peirce criterion. These preprocessing stages of SRQE-Boost model accurately organize the dataset into suitable format. In addition, Factor regressive analysis is used for measuring the relationships between features and the target based on Tanimoto Similarity Index. This regression function selects the significant features and removed the others, thereby reducing the time consumption of heart disease prediction.

**Table 11 Comparison of Memory Consumption**

| **Number of patient data** | **Memory consumption (KB) (without feature selection)** | | | **Memory consumption (KB) (with feature selection)** | | |
|---|---|---|---|---|---|---|
| | Proposed SRQE-Boost | LR [1] | stacking-based classification model [2] | Proposed SRQE-Boost | LR [1] | stacking-based classification model [2] |
| **7000** | 156.8 | 178.6 | 165.8 | 133.5 | 158.2 | 145.6 |
| **14000** | 162.2 | 192.6 | 175.5 | 142.5 | 175.6 | 158.4 |
| **21000** | 170.6 | 201.5 | 182.4 | 150.8 | 185.3 | 162.2 |
| **28000** | 175.8 | 216.6 | 192.6 | 154.7 | 190.2 | 170.5 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **35000** | 180.2 | 224.8 | 201.2 | 158.2 | 195.3 | 175.8 |
| **42000** | 184.7 | 232.6 | 212.6 | 164.5 | 205.2 | 181.2 |
| **49000** | 188.3 | 245.7 | 219.7 | 168.7 | 209.6 | 183.5 |
| **56000** | 192.6 | 252.3 | 222.6 | 175.2 | 213.3 | 187.6 |
| **63000** | 201.3 | 259.6 | 232.4 | 179.2 | 217.6 | 195.8 |
| **70000** | 212.5 | 262.6 | 238.2 | 188.5 | 220.6 | 206.2 |

Table 11 describes the performance analysis of the memory consumption with and without feature selection versus number of patient data. The numbers of data are taken in the ranges from 7000 to 70000. The memory consumption using the proposed SRQE-Boost model is considerably reduced than the other two existing feature selection methods. The overall results of the SRQE-Boost model were compared to the results observed by using existing classification methods [1] and [2]. The average of these ten comparison results illustrates that the memory consumption using the SRQE-Boost model with feature selection was considerably minimized by 18% and 9% compared to methods [1] and [2], respectively. In addition, the average of ten comparison results indicates that memory consumption using the SRQE-Boost model without feature selection was significantly reduced by 19% and 10% compared to methods [1] and [2], respectively. The two dimensional graphical analysis of memory consumption is shown in figure 11.

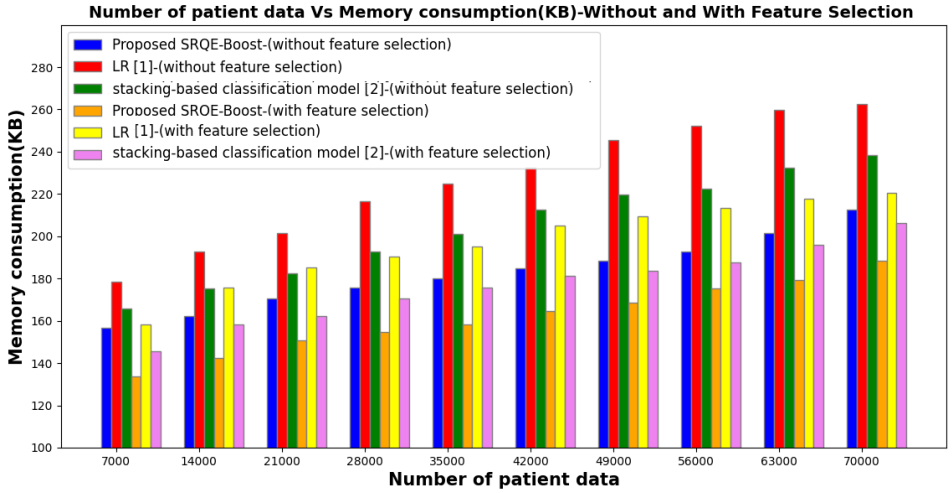



**Figure 16 Performance Analysis of Memory Consumption**

Figure 16 given above illustrates the graphical analysis of memory consumption with and without feature selection for disease prediction with respect to the number of patient data, ranging from 7000 to 70,000. As shown in graph, the memory consumption increases for all three methods as the number of patient data increases linearly. This improvement is achieved by removing the outlier data from the dataset using Peirce criterion through the mean and deviation analysis. In addition, the significant feature selection process of the SRQE-Boost model also removes the irrelevant features columns, thereby efficiently reduced the storage space.

**Table 12 Comparison of Feature Selection**

| Methods | Total number of features | Number of selected features |
|---|---|---|
| Proposed SRQE-Boost model | 13 | 8 |
| LR [1] | 13 | 10 |
| Stacking-based classification model [2] | 13 | 9 |

Table 12 illustrates the comparison analysis of number of features selected from the dataset by applying three different methods namely SRQE-Boost model, LR [1], stacking-based classification model [2]. From the tabulated results, SRQE-Boost model accurately selecting the eight most features as more relevant features such as age, gender, patient height, weight, cholesterol levels, systolic blood pressure, Diastolic blood pressure and glucose level.

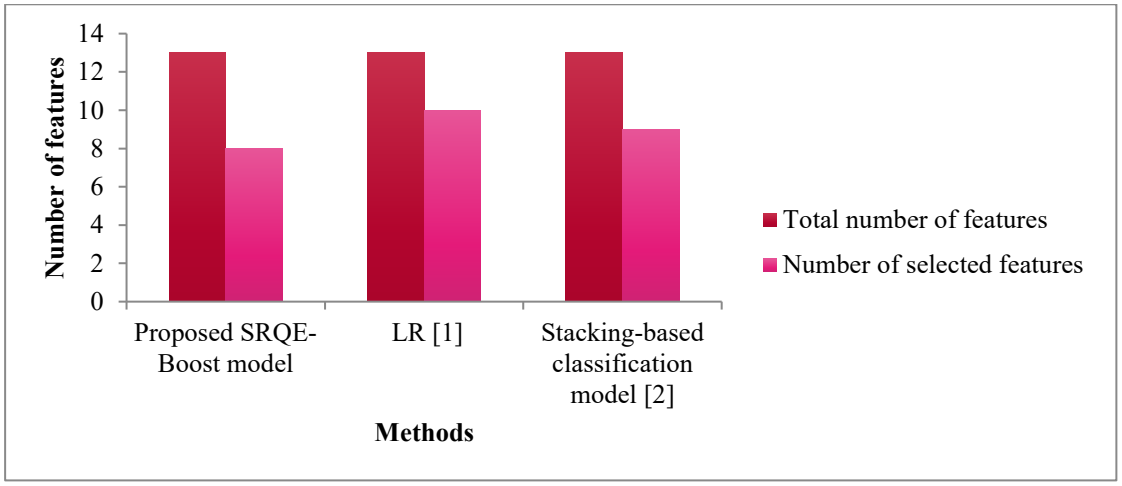



**Figure 17 Performance Analysis of Feature Selection**

Figure 17 illustrates the performance analysis of relevant feature selection using three different methods namely SRQE-Boost model, LR [1], stacking-based classification model [2]. From the observed experimental results, the SRQE-Boost model outperforms the other existing methods in the relevant feature selection process. As shown in the results, SRQE-Boost model selected 8 features from the dataset, while the existing methods [1] [2] selected 10 and 9 relevant features for heart disease prediction, respectively.

**Confusion matrix**

A confusion matrix serves as a critical evaluation tool in feature selection tasks, particularly for assessing the performance of the proposed SRQE-Boost model, LR [1], stacking-based classification model [2]. The matrix outlines four key components such as True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). By analyzing the confusion matrix, the strengths and weaknesses of the proposed SRQE-Boost model, LR [1], stacking-based classification model [2] clearly identified.

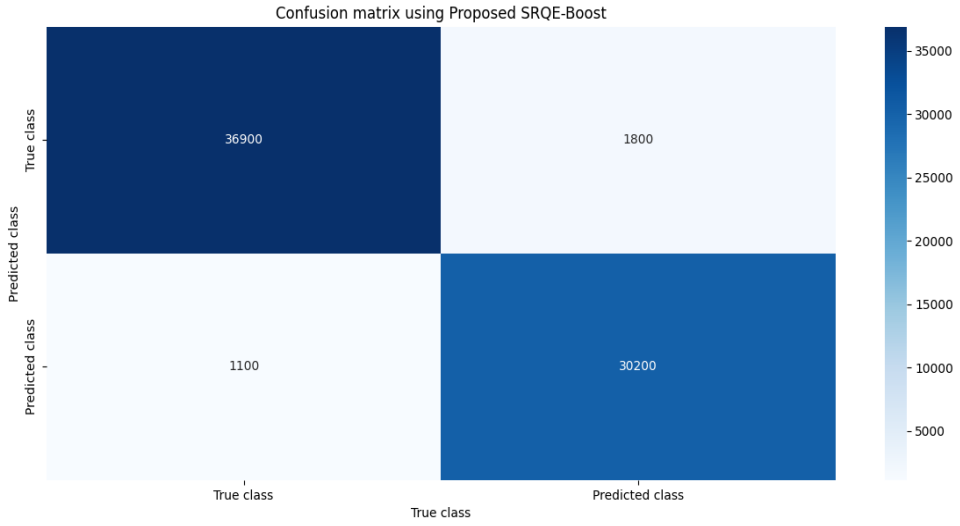



**Figure 18 Confusion Metrics Using Proposed SRQE-Boost model**

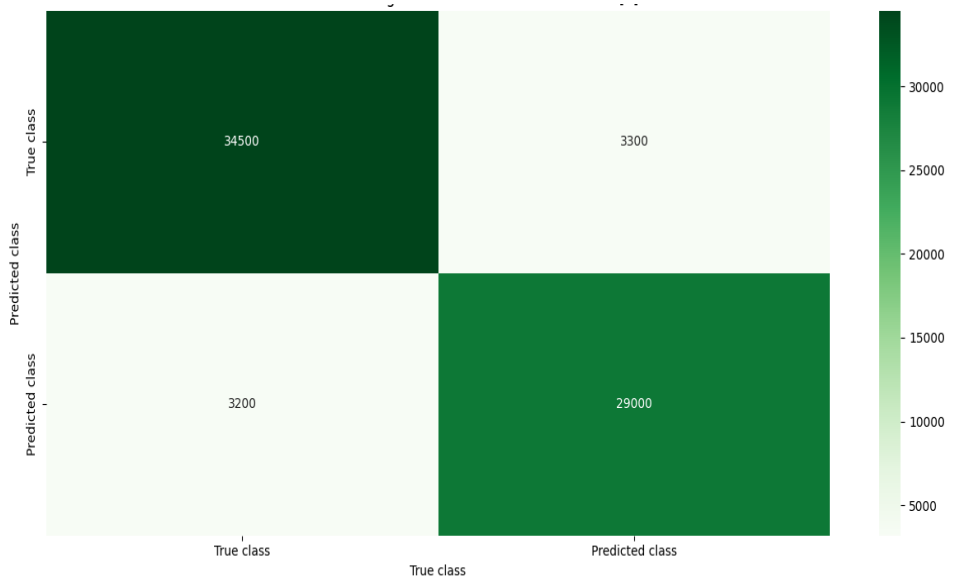



**Figure 19 Confusion Metrics Using LR**

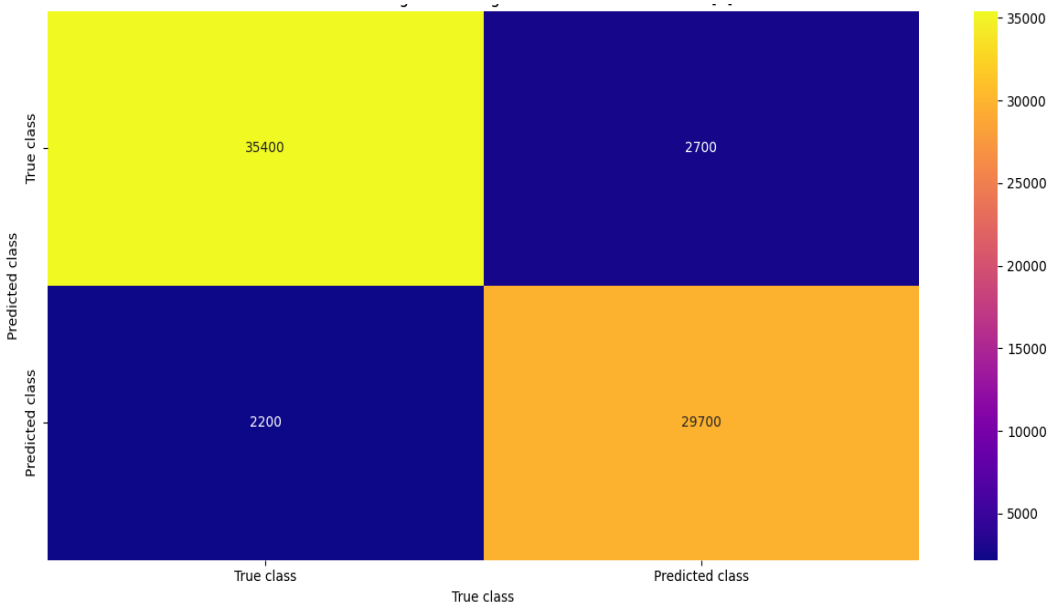



**Figure 20 Confusion Metrics Using Stacking-based Classification model [2].**

Figure 18, 19, 20 presents the confusion matrices generated by three different models SRQE-Boost model, LR [1], stacking-based classification model [2]. These matrices provide a visual representation of how effectively each model predicts heart disease risk levels using a dataset of 70,000 samples.

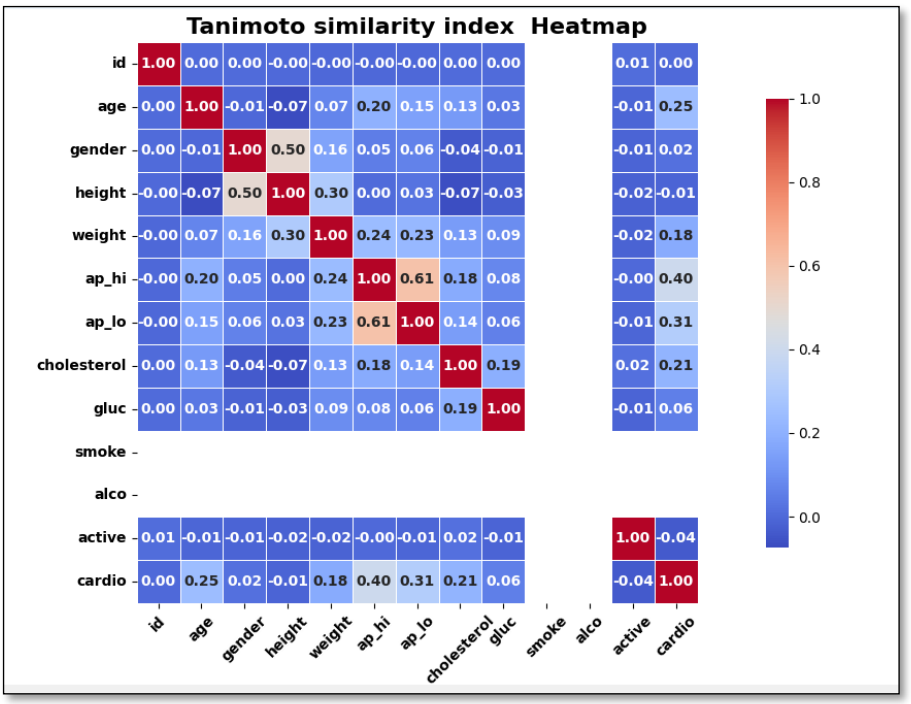



**Figure 21 Tanimoto Similarity Coefficients Between All the Features for Original Dataset**

Figure 21 illustrates the Tanimoto similarity coefficients computed for all the features within the dataset. The Tanimoto coefficient measures the similarity between two sets, commonly used to evaluate the redundancy or overlap among selected features. In the context of feature selection, a lower similarity score indicates a more diverse and less redundant set of features, which leads to better generalization and improved model performance. The proposed method, compared to existing techniques, demonstrates an optimal balance by minimizing redundant features while retaining more relevant features.

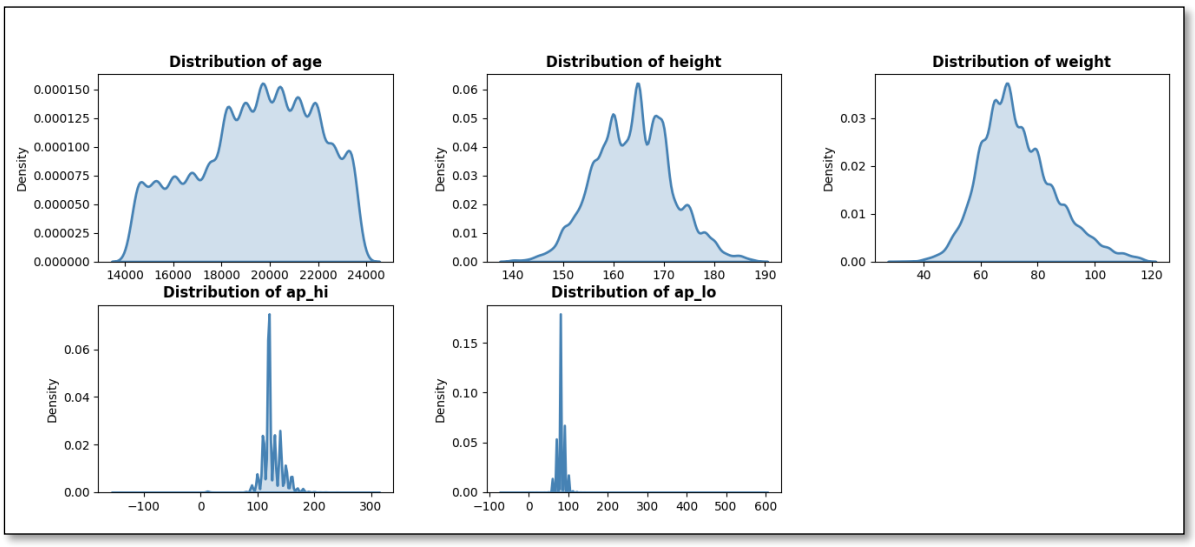



**Figure 22 distributions of numerical features**

Figure 22 illustrates the distribution patterns of numerical features in the given cardiovascular disease dataset. The distribution graph helps for understanding the range, mean, and deviation of each feature, as well as identifying potential outliers, skewness, or imbalances in the data. This analysis is essential for data preprocessing and influence the overall performance of the machine learning model.

## 6. CONCLUSION

Heart diseases are a leading cause of death worldwide, making early detection essential for improving the patient health. In this paper, the SRQE-Boost model is designed for heart disease prediction. The SRQE-Boost model integrates key steps such as pre-processing and feature selection and classification. Through the data pre-processing of the SRQE-Boost model reduces the time as well as memory consumption. Additionally, the feature selection process, utilizing the Tanimoto indexive factor regressive analysis, enhances the feature selection to further reduce the time consumption. Finally, the Quadratic Discriminant Emphasis Boosting ensemble classifier increases the accuracy of disease prediction. A comprehensive experimental evaluation was performed using various performance metrics, including prediction accuracy, precision, recall, F1 score, specificity, MCC, ROC-AUC, confusion matrix, prediction time and memory consumption across different patient data. The quantitative analysis indicates that the proposed SRQE-Boost model considerably enhances prediction accuracy while reducing time as well as memory consumption compared to existing approaches

## REFERENCES


1. G. Manikandan, B. Pragadeesh, V. Manojkumar, A.L. Karthikeyan, R. Manikandan, Amir H. Gandomi, “Classification models combined with Boruta feature selection for heart disease prediction”, Informatics in Medicine Unlocked, Elsevier, Volume 44, 2024, Pages 1-12. https://doi.org/10.1016/j.imu.2023.101442
2. Krishnamoorthy Natarajan, V. Vinoth Kumar, T. R. Mahesh, Mohamed Abbas, Nirmaladevi Kathamuthu, E. Mohan & Jonnakuti Rajkumar Annand, “Efficient Heart Disease Classification Through Stacked Ensemble with Optimized Firefly Feature Selection”, International Journal of Computational Intelligence Systems, Springer, Volume 17, 2024, Pages 1-14. https://doi.org/10.1007/s44196-024-00538-0
3. S. Venkatesh Babu, P. Ramya & Jeffin Gracewell, “Revolutionizing heart disease prediction with quantum-enhanced machine learning”, Scientific Reports, Volume 14, 2024, Pages 1-16. https://doi.org/10.1038/s41598-024-55991-w
4. K. Babu, A. Gokula Chandar & S. Kannadhasan, “Prediction and diagnosis of cardiovascular disease using cloud and machine learning design”, Journal of Cloud Computing, Springer, Volume 14, 2025, Pages 1-9. https://doi.org/10.1186/s13677-024-00720-x
5. Ibomoiye Domor Mienye and Nobert Jere, “Optimized Ensemble Learning Approach with Explainable AI for Improved Heart Disease Prediction”, Information, Volume 15, Issue 7, 2024, Pages 1-15. https://doi.org/10.3390/info15070394
6. G. Madhukar Rao, Dharavath Ramesh, Vandana Sharma, Anurag Sinha, Md. Mehedi Hassan & Amir H. Gandomi, “AttGRU-HMSI: enhancing heart disease diagnosis using hybrid deep learning approach”, Scientific Reports, Volume 14, 2024, Pages 1-19. https://doi.org/10.1038/s41598-024-56931-4
7. Najmu Nissa, Sanjay Jamwal and Mehdi Neshat, “A Technical Comparative Heart Disease Prediction Framework Using Boosting Ensemble Techniques”, Computation, Volume 12, Issue 1, 2024, Pages 1-22. https://doi.org/10.3390/computation12010015
8. Subhash Mondal, Ranjan Maity, Yachang Omo, Soumadip Ghosh, Amitava Nag, “An Efficient Computational Risk Prediction Model of Heart Diseases Based on Dual-Stage Stacked Machine Learning Approaches”, IEEE Access, Volume 12, 2024, Pages 7255 – 7270. DOI: 10.1109/ACCESS.2024.3350996
9. Pierre Claver Bizimana, Zuping Zhang, Alphonse Houssou Hounye, Muhammad Asim, Mohamed Hammad & Ahmed A. Abd El-Latif, “Automated heart disease prediction using improved explainable learning-based technique”, Neural Computing and Applications, Springer, Volume 36, 2024, Pages 16289–16318. https://doi.org/10.1007/s00521-024-09967-6
10. B. Ramesh, Kuruva Lakshmanna, “A Novel Early Detection and Prevention of Coronary Heart Disease Framework Using Hybrid Deep Learning Model and Neural Fuzzy Inference System”, IEEE Access, Volume 12, 2024, Pages 26683 – 26695. DOI: 10.1109/ACCESS.2024.3366537
11. E. I. Elsedimy, Sara M. M. AboHashish & Fahad Algarni, “New cardiovascular disease prediction approach using support vector machine and quantum-behaved particle swarm optimization”, Multimedia Tools and Applications, Springer, Volume 83, 2024, Pages 23901–23928. https://doi.org/10.1007/s11042-023-16194-z
12. Hosam El-Sofany, Belgacem Bouallegue & Yasser M. Abd El-Latif, “A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method”, Scientific Reports, Volume 14, 2024, Pages 1-18. https://doi.org/10.1038/s41598-024-74656-2
13. Amrit Singh, Harisankar Mahapatra, Anil Kumar Biswal, Madhumita Mahapatrad, Debabrata Singh, Milan Samantaray, “Heart Disease Detection Using Machine Learning Models”, Procedia Computer Science, Elsevier, Volume 235, 2024, Pages 937-947. https://doi.org/10.1016/j.procs.2024.04.089
14. Awad Bin Naeem, Biswaranjan Senapati, Dipen Bhuva, Abdelhamid Zaidi, Abhishek Bhuva, Md. Sakiul Islam Sudman, “Heart Disease Detection Using Feature Extraction and Artificial Neural Networks: A Sensor-Based

Approach", IEEE Access, Volume 12, 2024, Pages 37349 – 37362. DOI: 10.1109/ACCESS.2024.3373646

15. Atta Ur Rahman, Yousef Alsenani, Adeel Zafar, Kalim Ullah, Khaled Rabie & Thokozani Shongwe, "Enhancing heart disease prediction using a self-attention-based transformer model", Scientific Reports, Volume 14, 2024, Pages 1-13. https://doi.org/10.1038/s41598-024-51184-7
16. Navita, Pooja Mittal, Yogesh Kumar Sharma, Umesh Kumar Lilhore, Sarita Simaiya, Kashif Saleem & Ehab Seif Ghith, "Advanced Hybrid Machine Learning Model for Accurate Detection of Cardiovascular Disease", International Journal of Computational Intelligence Systems, Springer, Volume 18, 2025, Pages 1-20. https://doi.org/10.1007/s44196-025-00771-1
17. Tahseen Ullah, Syed Irfan Ullah, Khalil Ullah, Muhammad Ishaq, Ahmad Khan, Yazeed Yasin Ghadi, "Machine Learning-Based Cardiovascular Disease Detection Using Optimal Feature Selection", IEEE Access, Volume 12, 2024, Pages 16431 – 16446. DOI: 10.1109/ACCESS.2024.3359910
18. Ayad E. Korial, Ivan Isho Gorial and Amjad J. Humaidi, "An Improved Ensemble-Based Cardiovascular Disease Detection System with Chi-Square Feature Selection", Computers, Volume 13, Issue 6, 2024, Pages 1-17. https://doi.org/10.3390/computers13060126
19. Shanshan Wang, Lei Zhang, Xiao Liu, Jiuye Sun, "Optimization of multidimensional feature engineering and data partitioning strategies in heart disease prediction models', Alexandria Engineering Journal, Elsevier, Volume 107, November 2024, Pages 932-949. https://doi.org/10.1016/j.aej.2024.09.037
20. Omar Mahmood Yaseen, Mohanad Mohammed Rashid, "An Explainable Artificial Intelligence (XAI) Methodology for Heart Disease Classification", International Journal of Current Science Research and Review, Volume 08, Issue 02, 2025, Pages 809-817. https://doi.org/10.47191/ijcsrr/V8-i2-28
21. Mohammed Amine Bouqentar, Oumaima Terrada, Soufiane Hamida, Shawki Saleh, Driss Lamrani, Bouchaib Cherradi, Abdelhadi Raihani, "Early heart disease prediction using feature engineering and machine learning algorithms", Heliyon, Elsevier, Volume 10, Issue 19, 2024, Pages 1-23. https://doi.org/10.1016/j.heliyon.2024.e38731
22. Temidayo Oluwatosin Omotehinwa, David Opeoluwa Oyewola, Ervin Gubin Moung, "Optimizing the light gradient-boosting machine algorithm for an efficient early detection of coronary heart disease", Informatics and Health, Elsevier, Volume 1, Issue 2, September 2024, Pages 70-81. https://doi.org/10.1016/j.infoh.2024.06.001
23. Francisco Mesquita, Gonçalo Marques, "An explainable machine learning approach for automated medical decision support of heart disease', Data & Knowledge Engineering, Elsevier, Volume 153, September 2024, pages 1-15. https://doi.org/10.1016/j.datak.2024.102339
24. Stephen Akatore Atimbire, Justice Kwame Appati & Ebenezer Owusu, "Empirical exploration of whale optimisation algorithm for heart disease prediction", Scientific Reports, volume 14, 2024, Pages 1-22. https://doi.org/10.1038/s41598-024-54990-1
25. Adedayo Ogunpola, Faisal Saeed, Shadi Basurra, Abdullah M. Albarrak and Sultan Noman Qasem, "Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases", Diagnostics, Volume 14, Issue 2, 2024, Pages 1-19. https://doi.org/10.3390/diagnostics14020144
26. Md. Liakot Ali, Muhammad Sheikh Sadi, Md. Osman Goni, "Diagnosis of heart diseases: A fuzzy-logic-based approach", PLoS ONE, Volume 19, Issue 2, 2024, Pages 1-25. https://doi.org/10.1371/journal.pone.0293112
27. Abhigya Mahajan, Baijnath Kaushik, Mohammad Khalid Imam Rahmani, Abdulbasid S. Banga, "A Hybrid Feature Selection and Ensemble Stacked Learning Models on Multi-Variant CVD Datasets for Effective Classification", DOI: 10.1109/ACCESS.2024.3412077
28. Syed Ali Jafar Zaidi, Attia Ghafoor, Jun Kim, Zeeshan Abbas and Seung Won Lee, "HeartEnsembleNet: An Innovative Hybrid Ensemble Learning Approach for Cardiovascular Risk Prediction", Healthcare, Volume 13, Issue 5, Pages 1-21. https://doi.org/10.3390/healthcare13050507
29. Sharanya Prabhu, Shourya Gupta, Gautham Manuru Prabhu, Aarushi Vishal Dhanuka, K. Vivekananda Bhat, "QuCardio: Application of Quantum Machine Learning for Detection of Cardiovascular Diseases", IEEE Access, Volume 11, 2023, Pages 136122 – 136135. DOI: 10.1109/ACCESS.2023.3338145
30. Muhammad Talha Ashfaq, Nadeem Javaid, Nabil Alrajeh & Syed Saqib Ali, "An explainable AI based new deep learning solution for efficient heart disease prediction at early stages", Evolving Systems, Springer, Volume 16, 2025, Pages 1-26. https://doi.org/10.1007/s12530-025-09664-2.