INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Big Data Analytics for Heart Disease Prediction using Regularized Principal and Quadratic Entropy Boosting

P Muthulakshmi[1]*, **M Parveen**[2]

1 Associate Professor, Cauvery College for Women (Autonomous), [Affiliated to Bharathidasan University], Tiruchirappalli, 620 018, Tamil Nadu, India
2 Professor, Cauvery College for Women (Autonomous), [Affiliated to Bharathidasan University], Tiruchirappalli, 620 018, Tamil Nadu, India

OPEN ACCESS

Received: 18-11-2023

Accepted: 02-01-2024

Published: 02-02-2024

Citation: Muthulakshmi P, Parveen M (2024) Big Data Analytics for Heart Disease Prediction using Regularized Principal and Quadratic Entropy Boosting. Indian Journal of Science and Technology 17(6): 533-547. https://doi.org/10.17485/IJST/v17i6.2928

* **Corresponding author**.

muthulakshmi.cs@cauverycollege.ac.in

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: Over the past few years, there prevails an abundance wealth of big data obtained via patients' electronic health records. One of the leading causes of mortality globally is the cardiovascular disease. Based on the present test and history cardiovascular disease diagnosing of patients can be done. Therefore, early and quick diagnosis can reduce the mortality rate. To address their needs, several machine learning methods have been employed in the recent past in cardiovascular disease diagnosis and prediction. Previous research was also concentrated on acquiring the significant features to heart disease prediction however less importance was given to the time involved and error rate to identifying the strength of these features. **Methods**: In this work we plan to develop a method called, Regularized Principal Component and Quadratic Weighted Entropy Boosting (RPC-QWEB) for predicting heart disease. Initially in RPC-QWEB, relevant features are selected to avoid missing values in the input database by employing Regularized Principal Component Regressive Feature Selection (RPCRFS). Second, with the obtained dimensionality reduced features, Quadratic Weighted Entropy Boosting Classification (QWEBC) process is carried out to classify the patient data as normal or abnormal. The QWEBC process is an ensemble of several weak classifiers (i.e., Quadratic Classifier). The weak classifier results are combined to form strong classifier and provide final prediction results as normal or abnormal condition with minimal error rate. **Findings**: Experimental evaluation is carried out on factors with the cardiovascular disease dataset such as heart disease prediction accuracy, heart disease prediction time, sensitivity, error rate with respect to distinct numbers of patient data. The proposed RPC-QWEB method was compared with existing Heart Disease Prediction Framework (HDPF) and Swarm Artificial Neural Network (Swarm-ANN). **Novelty**: RPC-QWEB method outperforms the conventional learning methods in terms of numerous performance matrices. The RPC-QWEB method produces 3% and 5% increase in terms of accuracy and sensitivity and 7% and 29% reduced prediction time and error rate as compared to the existing benchmark methods. We may use this method to predict the heart disease at early stage there by we can reduce the death rate.

Keywords: Big data; Regularized Principal Component; Quadratic Weighted Entropy Boosting; Regressive Feature Selection; Classification

1 Introduction

Big data is a high-volume and contain variety of information. Big data plays an essential part in heart disease prediction. Over the past few years, one of the major causes of mortality globally is cardiovascular or heart disease. World Health Organization (WHO) has investigated that more than 18 million deaths persist in a year on an average in the world owing to cardiovascular disease. Also, one of the major causes for death is because of heart disease that comprises heart attack, hypertension, and stroke. In such a circumstance, cardiac disease early diagnosis can assist in taking stern action with accurate and robust treatment therefore circumventing death. Due to this, heart disease prediction at early stage on the basis of monitoring dataset in a remote fashion is considered to be challenging task as far as healthcare domain is concerned.

The proposed RPC-QWEB method is performed to improve the heart disease accuracy rate. RPC Regressive Feature Selection is applied for analyzing the linear regression to gather the significant features for efficient heart disease prediction. In order to reduce the heart disease diagnosis time. Novelty of QWEBC is applied for patient data classification. RPC-QWEB method provides better performance in terms of accuracy, prediction time, error rate and sensitivity.

An intelligent healthcare method for cardiovascular heart disease prediction based on Swarm-Artificial Neural Network (Swarm-ANN) was proposed in[1]. To start with, the Swarm-ANN generated predefined numbers of Neural Networks (NNs) for training the network. Followed by which on the basis of their consistency in the acquired solution, the framework was further evaluated. Moreover, the NN populations were trained by utilizing two weights by updating it on the basis of heuristic formulation.

Finally, the neuron weight was modified by exchanging global best weight with other neurons with which the accuracy rate of cardiovascular disease prediction was made. Despite improvement observed in accuracy, the time involved in the prediction process was not focused. To address this issue, the proposed work employs Regularized Principal Component Regressive Feature Selection model that by applying principal component regression function, relevant features in a dimensionality reduced manner is obtained. By applying this dimensionality reduced features in the further classification stage, time involved in heart disease prediction can be improved significantly.

Supervised machine learning algorithms are significant classification mechanisms frequently utilized in constructing prediction models that assists in disease diagnosing at an early stage. However, certain issues like, overfitting and underfitting required to be addressed while constructing the mechanism. Heart Disease Prediction Framework (HDPF) or hybrid classifiers employing the ensemble model with a majority voting mechanism to enhance prediction accuracy was proposed in[2].

Moreover, pre-processing and features selection employing genetic algorithm was also designed with the purpose of improving prediction performance and overall time consumption. Despite improvement observed both in terms of performance and time the error rate involve during heart disease prediction was not focused. To reduce the error rate involving heart disease prediction, Quadratic Weighted Entropy Boosting Classification is applied that boosts the results of the weak learners into strong learners, therefore enhancing the overall classification rate. With the improvement in classification rate, the error involved in heart disease prediction can be reduced to a greater extent.

In[3], a novel method with the objective of identifying relevant features by means of machine learning techniques for predicting cardiovascular disease was presented with which accurate results were arrived at. Here, the prediction was made by means of several mixtures of features and classification patterns, therefore enhancing the accuracy rate. However, it becomes a laborious and tedious process to select significant features from heterogeneous electronic health records and also poses severe threat to acquire precise characterizations for patients.

In [4], Deep Risk based on attention mechanism and deep neural networks was proposed. With this integrated mechanisms, not only robust features were said to be learnt in an automatic manner, but also predicted patients' risk of cardiovascular diseases in a timely manner. An efficient neural network with convolutional layers was introduced in [5] to categorize class-imbalanced clinical data. The data was acquired from National Health and Nutritional Examination Survey (NHANES) for predicting Coronary Heart Disease (CHD) occurrence.

Also, a least absolute shrinkage and selection operator (LASSO) based feature weight assessment model was introduced with the identification of feature based on the majority-voting. Significant features were homogenized by fully connected layer before sending the output to consecutive convolutional stages. Moreover, a simulated annealing process was also utilized for increasing the classification accuracy. However, CNN was not implemented for similar clinical dataset prediction where imbalanced number of positive and negative classifications was said to exist.

An atherosclerotic cardiovascular disease (ASCVD) risk prediction model was introduced in [6] for patients with preceding ASCVD on statin use. The statin-treated patients were used for simulation with ASCVD from AIM-HIGH trial cohort. Also, a 5-year risk score was computed for successive ASCVD events using Cox regression employing potential risk factors like age, sex, and race. However, the computational cost was not minimized by ASCVD risk prediction model.

Predictive value of pathological factors was determined in [7] for early Heart Failure detection via social network based approach. Here, the electronic health records were gathered with the objective of determining the similarity of risk factors. The similarity values were utilized to construct both weighted and unweighted medical social network. On one hand, the constructed medical social network was split into two types of risk groups namely, HF high-risk group and HF low-risk group by means of group division algorithm. With this despite the improvement of accuracy, the cost and time involved in the process was not improved by designed approach.

A machine learning algorithm was introduced in [8] for prediction using training data. With the information provided by the user the prediction result was obtained. However, the accuracy rate was found to decrease with the incomplete medical data. To address on this aspect, convolutional neural network algorithm was introduced with structured and unstructured patient data. However, the prediction time was not reduced by designed machine learning algorithm.

Artificial intelligence can aid the providers in different types of patient care and intelligent health patterns. As far as artificial intelligence methods are concerned, both machine learning and deep learning are found in assisting disease diagnosis, discovery of different types of drug and identification of risk involved.

A comprehensive survey was designed in [9] on the basis of artificial intelligence techniques for diagnosing several diseases like, Alzheimer, diabetes, heart disease, stroke, tuberculosis, and so on. Yet another smart healthcare framework was proposed in [10] on the basis of deep and machine learning using optimization stochastic gradient descent. In this method good accuracy was attained.

Immense studies provide only a glance into heart disease prediction using machine learning techniques. In [3], a novel method was designed with the objective of identifying pertinent features by means of machine learning techniques, therefore enhancing the accuracy rate significantly in cardiovascular disease prediction. Here, the prediction was introduced by utilizing several mixtures of features, and numerous classification models, therefore contributing to accuracy significantly.

With the objective of analyzing the present data in predicting results in an optimal manner, optimization techniques are required. In [11], a framework for predicting heart disease utilizing several risk factors on the basis of numerous classification algorithms was proposed. Moreover, the highest performance was attained using Bayesian Optimized Support Vector Machine.

Early heart disease prediction not only assists the patients circumvent it, but also assists the medical professionals identify the major reasons of heart attack and prevent it prior to its actual happening. In [12], Cardio Help was proposed that predicting the probability of presence of cardiovascular disease by means of deep learning algorithm called convolution neural networks (CNN). Also, temporal data was also involved for heart failure prediction at its earliest stage therefore improving accuracy.

Prediction of cardiac disease assists practitioners make decisions accurately concerning patients' health. Hence, the utilization of machine learning (ML) is said to be one of the solutions in minimizing and apprehending heart disease symptoms. In spite of the ceaseless development of medical practices, diseases related to heart still remains the major cause of death.

In [13], a novel method was proposed with the objective of extracting features related to Electro Cardio Grams. With this feature type, real time analysis was made for heart disease prediction. However, medical practitioners frequently diagnose cardiovascular disease on the basis of the present and past clinical test results for diagnosing patients with indistinguishable indications. This is owing to the reason that the patients suffering from heart disease be in need of swift diagnosis, prompt therapy and ceaseless considerations. To address on these aspects, several data mining methods have been employed at one time with the purpose of diagnosing and predicting heart diseases at an early stage, however less significance was provided to recognizing the feature strengths. In[14] weighted association rule mining was applied to the dataset for acquiring significant features. With the acquired significant features, heart disease prediction was made in a timely and accurate manner.

Heart disease prediction assists the physicians or medical practitioners in making more precise and robust decisions concerning patients' health. Hence, the utilization of machine learning (ML) is found to be a solution in minimizing and apprehending the symptoms pertaining to heart disease. A dimensionality reduction method by employing machine learning technique to identify significant features of heart disease via feature selection model was proposed in[15]. With this feature selection model highest accuracy was said to be obtained.

In[16], an algorithm called sequential feature selection was designed with the purpose of enhancing the performance involving classification on detecting heart disease. This was performed by eliminating features found to be irrelevant and of the least significance and conducting training employing optimal features. With this the accuracy rate was said to be improved extensively. Despite improvement observed in accuracy, the error rate was not focused.

Yet another decision tree based random forest model was proposed in[17] with minimum error rate. Another method concentrating of feature selection aspect was designed in[18] by employing machine learning methods therefore resulting in better performance. An elaborate review on heart disease prediction employing data mining and machine learning were investigated in[19].

A heart disease prediction algorithm was designed in[20] which combine the embedded feature selection method and deep neural networks. But heart disease prediction accuracy was not improved. A smart healthcare system was proposed in[21] for heart disease prediction using ensemble deep learning and feature fusion approaches. However, heart disease prediction time was not reduced.

A hybrid one-dimensional convolutional neural network (1D CNN) was proposed in[22] which uses a large dataset accumulated from online survey data and selected features using feature selection algorithms. The non-coronary heart disease (no-CHD) and CHD validation data showed an accuracy of 80.1% and 76.9%, respectively. The model was compared with an artificial neural network, random forest, AdaBoost, and a support vector machine. Overall, 1D CNN proved to show better performance in terms of accuracy, false negative rates, and false positive rates but failed to handle imbalanced data.

In order to overcome the above issues a method, called, Regularized Principal Component and Quadratic Weighted Entropy Boosting (RPC-QWEB) is introduced for predicting heart disease. The main objective of this work is to improve the performance accuracy of heart disease prediction in a timely and minimum error rate. Many works have been proposed with significant features. In contrast, the RPC-QWEB method uses all features without any restrictions of feature selection and by means of regularized principal components acquire the significant features for further processing. Ensemble model is applied to the selected significant features for classification between normalcy and abnormality of patient data. The elaborate description of the proposed RPC-QWEB method is presented in the following sections.

1.1 Major Contributions

To overcome the issues from the literature review, an RPC-QWEB method is introduced with the novel contributions as given below,

- To improve heart disease diagnosis accuracy rate, an RPC-QWEB method is introduced on the basis of two distinct processes namely, feature selection and classification
- To minimizing heart disease diagnosis time and improve accuracy rate, Regularized Principal Component Regressive Feature Selection is applied for analyzing the linear regression and to select the significant features for efficient heart disease prediction
- Quadratic Weighted Entropy Boosting Classification (QWEBC) is then applied for patient data classification. The QWEBC algorithm initially utilizes a Quadratic Likelihood Ratio classifier to classify the patient data based on the resulting separating surface between classes is quadratic between testing and training sample. The Weighted Entropy Boosting combines weak classifier performance to make strong results by reducing the generalization error. This helps to reduce the error rate involved in heart disease prediction
- An immense experiment is organized to measure the performance of the RPC-QWEB method and state-of-the-art works. The results achieved shows that our proposed, RPC-QWEB method provides better performance in terms of accuracy, time, error rate and sensitivity

1.2 Organization of the Paper

The rest of the paper is organized as follows. Section 2 provides the methodology. Regularized Principal Component and Quadratic Weighted Entropy Boosting method by separating into two sections, feature selection and classification modelling. Section 3 provides the experimental setup, with the detailed discussion about the RPC-QWEB method results and benchmarking of the proposed method. Finally, Section 4 ends with a conclusion of current work.

2 Methodology

2.1 Regularized Principal Component and Quadratic Weighted Entropy Boosting

A Regularized Principal Component and Quadratic Weighted Entropy Boosting (RPC-QWEB) method is designed for predicting heart disease in a timely and accurate manner with minimum error rate. The RPC-QWEB method is split into two sections. They are feature selection and classification. Feature selection is performed by means of Regularized Principal Component Regression function. Second classification is performed by applying Quadratic Weighted Entropy Boosting Classification to the dimensionality reduced features. With these two processes, heart disease prediction is made in a computationally efficient manner. The elaborate description of the proposed RPC-QWEB method is provided in the following sections. Figure 1 given below shows the block diagram of RPC-QWEB method.

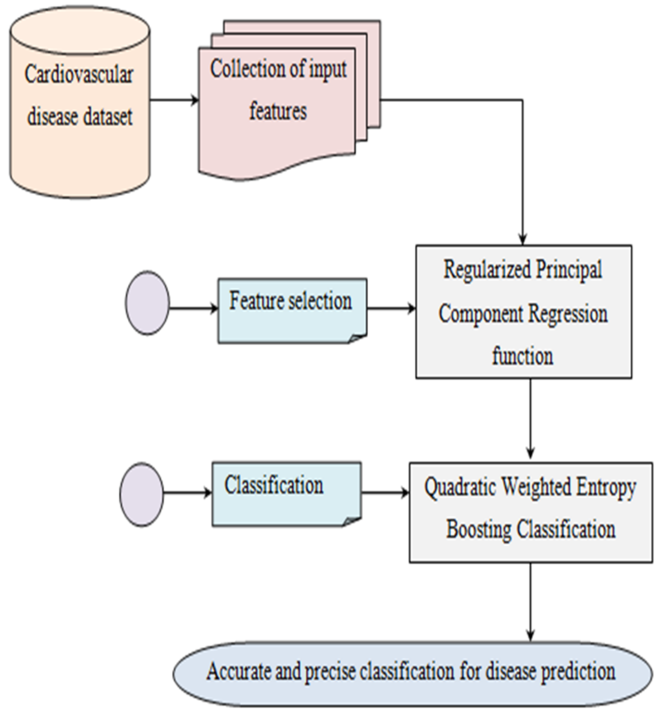
Cardiovascular disease dataset
Collection of input features
Regularized Principal Component Regression function
Feature selection
Classification
Quadratic Weighted Entropy Boosting Classification
Accurate and precise classification for disease prediction

Fig 1. Block diagram of RPC-QWEB

As shown in the figure, with the cardiovascular disease dataset provided as input, first dimensionality reduced and relevant features are selected using the Regularized Principal Component Regression function. Next, with the relevant features obtained, Quadratic Weighted Entropy Boosting Classification is performed to obtain accurate and precise classification results with minimum error rate. The elaborate description of the RPC-QWEB method is provided in the following sections.

The cardiovascular disease dataset considered in our work had irrelevant or repeated features. Table 1 shows three types of input features (i.e., objective – factual information, examination – results of medical examination and subjective – information given by the patient) included in the analysis. For this investigation, cardiovascular disease dataset comprising 12 features is considered.

Table 1. Cardiovascular dataset details

S. No	Features Or Data	Types Of Feature
1	Age	Objective feature
2	Height	Objective feature
3	Weight	Objective feature
4	Gender	Objective feature
5	Systolic blood pressure	Examination feature
6	Diastolic blood pressure	Examination feature
7	Cholesterol	Examination feature
8	Glucose	Examination feature
9	Smoking	Subjective feature
10	Alcohol intake	Subjective feature
11	Physical activity	Subjective feature
12	Presence or absence of cardiovascular disease	Target variable

With the above system model, the proposed RPC-QWEB method is designed involving feature selection and classification for early heart disease prediction.

2.2 Regularized Principal Component Regressive Feature Selection

Initially in RPC-QWEB method, the feature selection is carried out to avoid missing values in the input database. The most important considerations for selecting relevant features are to allocate a single category for missing values. For example a patient record value cannot have cholesterol or age equal to zero. If this occurs, the value of the corresponding will be swapped to the ‘*null value*’ class. With this initial consideration a Regularized Principal Component Regressive Feature Selection model is applied to the input dataset with the purpose of selecting the relevant features via dimensionality reduction.

Dimensionality reduction refers to the process of minimizing the number of features in concern with the purpose of extracting latent features for missing values from raw datasets while retaining the organization. The RPCRFS being a regression analysis determines the unknown regression coefficients in linear regression model. With this utilization of regression coefficients, dimensionality reduction is said to be achieved employing minimum number of parameters. Figure 2 shows the block diagram of Regularized Principal Component Regressive Feature Selection model.

Given the patient records from a dataset ‘*DS*’, the vector space is mathematically stated as given below.

$$VS = \begin{bmatrix} R_1F_1 & R_1F_2 & R_1F_3 & R_1F_4 & \dots & R_1F_n \\ R_2F_2 & R_2F_2 & R_2F_3 & R_2F_4 & \dots & R_2F_n \\ \dots & \dots & \dots & \dots & \dots & \dots \\ R_nF_1 & R_nF_2 & R_nF_3 & R_nF_4 & \dots & R_nF_n \end{bmatrix} \tag{1}$$

From the above Equation (1), the vector space ‘VS’ denotes the linear representation of each record ‘R_i’ with respect to the corresponding features ‘F_j’. Then, with the given patient records from a dataset ‘*DS*’, the principal components of ‘*k*’ is defined via orthonormal basis to obtain latent features is acquired as a solution as stated below.

$$LF = argmin \sum_{i=1}^{n} \left| F_i - \sum_{j=1}^{m} < F_i, \alpha_j > \right|_g^2 \tag{2}$$

From the above Equation (2), with the actual features ‘F_i’ of concern acquired from the cardiovascular dataset ‘*DS*’ provided as input, regularized features linearly with respect to the regression is formulated in such a manner so as to reduce the dimension. Hence, the training set features ‘α_j’ with the objective of tuning the entire set of records is first subjected to the global ‘*g*’ minimum to the constraints (i.e., acquiring relevant features).

Next, with the assumption that the functions possess a finite gradient dimension ‘$\left|\beta_g F\right|_g^2$’ on the entire patient record set, the linear regression coefficients are then formulated via eigen functions as a minimization of the Dirichlet energy.

$$DE = argmin \sum_{j=1}^{m} \left|\beta_g \alpha_j\right|_g^2, \ such\ that < \alpha_i, \alpha_j >_g = \gamma_{ij} \tag{3}$$

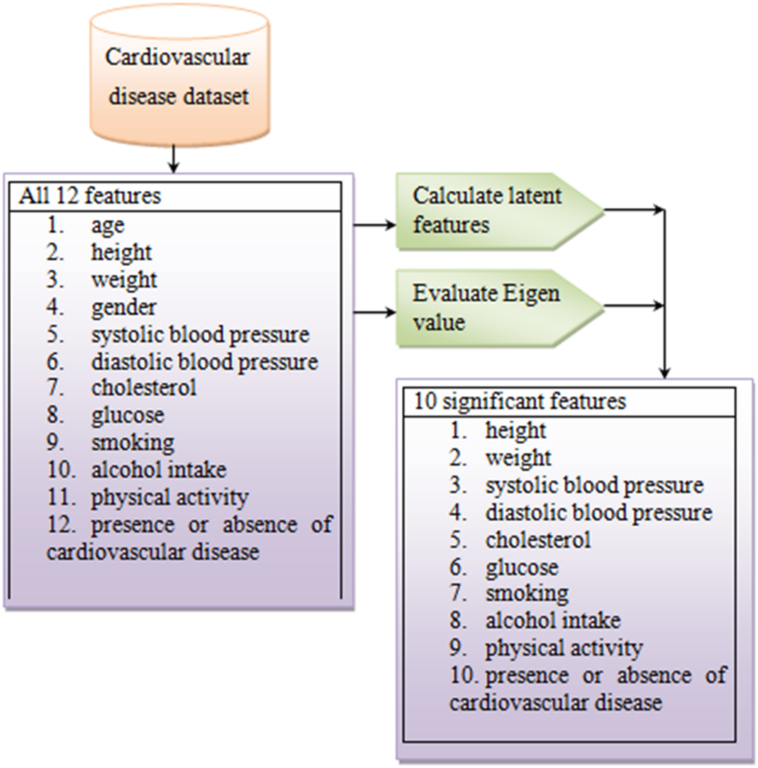
Cardiovascular disease dataset
All 12 features
1. age
2. height
3. weight
4. gender
5. systolic blood pressure
6. diastolic blood pressure
7. cholesterol
8. glucose
9. smoking
10. alcohol intake
11. physical activity
12. presence or absence of cardiovascular disease
Calculate latent features
Evaluate Eigen value
10 significant features
1. height
2. weight
3. systolic blood pressure
4. diastolic blood pressure
5. cholesterol
6. glucose
7. smoking
8. alcohol intake
9. physical activity
10. presence or absence of cardiovascular disease

Fig 2. Block diagram of Regularized Principal Component Regressive Feature Selection

From the above Equation (3), 'g_{ij}' represent the mapping between an arc length on a given record set space (i.e., denoting three types of input features) to length on the given manifold (i.e., with respect to each feature) to select relevant and significant features. Then, on the record set 'DS', the geometric projection '$Proj$' of 'F_i' is mathematically stated as given below.

$$argmin \sum_{i=1}^{n} \left| Proj\ Proj^T Weight\ [F_i] - F_i \right|_g^2 \tag{4}$$

From the above Equation (4), geometric projection '$Proj$' of 'F_i' via individual feature weights '$Weight\ (F_i]$' is formulated. Finally, to determine the amount of significant and relevant features to be selected, eigen value-one criterion is employed for analysis. This is mathematically stated as given below.

$$VS(F) = \gamma(F) \tag{5}$$

With this, all the features with an eigen value 'γ' greater than '1' are retained. Hence, the eigen feature values greater than '1' perch for higher contrast than their contribution as actual variables. On contrary, eigen feature value less than '1' contributed less than their actual value and were eliminated or removed from further analysis. Therefore, the features with eigen feature values greater than '1' are retained and considered as relevant features whereas the features with eigen feature values less than '1' are considered as irrelevant features and removed from further processing. The pseudo code representation of Regularized Principal Component Regressive Feature Selection is given below:

Algorithm 1. Regularized Principal Component Regressive Feature Selection

Input: Dataset 'DS', Feature '$F = F_1, F_2, \ldots, F_n$', Feature type '$T = 1, 2, 3$', Feature type 1 '$T_1 = O$', Feature type 2 '$T_3 = E$', Feature type 3 '$T_3 = S$'

Output: Computationally efficient robust and accurate feature selection

```
Step 1: Initialize record '$R_i$', training set features '$\alpha_j$', Weight '$Weight$'
Step 2: Begin
Step 3: For each Dataset '$DS$' with Feature '$F$'
Step 4: Mathematically formulate the vector space as in Equation (1)
Step 5: Obtain latent features as in Equation (2)
Step 6: Formulate linear regression coefficients as in Equation (3)
Step 7: Model geometric projection as in Equation (4)
Step 8: Evaluate eigen value-one criterion as in Equation (5)
Step 9: If eigenvalue '$\gamma > 1$'
Step 10: Features are retained
Step 11: Selected features are '$FS$'
Step 12: End if
Step 13: If eigen value '$\gamma \leq 1$'
Sep 14: Features are removed
Step 15: End if
Step 16: End for
Step 17: End
```

As given in the above algorithm, the first step is to estimate the individual feature weights. Let '$F = F_1,\ F_2,\ \ldots, F_n$' denote the set of features and these features are stored in the form of vector. Then, with the features represented in vector space, latent features are obtained. Followed by which, geometric projection for each latent features are acquired. Next, the eigen value-one criterion is evaluated with which the eigen value is measured by means of conditionality checking. Finally, with the total number of features in the cardiovascular dataset being 12 after feature selection, it is reduced to 10 (i.e., height, weight, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoking, alcohol intake, physical activity and presence or absence of cardiovascular disease). As a result, the relevant features are selected in a computationally efficient and accurate manner.

2.3 Quadratic Weighted Entropy Boosting Classification

With the dimensionality reduced features, Quadratic Weighted Entropy Boosting Classification (QWEBC) is performed to classify the patient data as normal or abnormal. The Quadratic Weighted Entropy Boosting is an ensemble classifier that converts weak learners into strong one. The weak learner (i.e., Quadratic Classifier) being a base classifier is said to be in dearth of providing precise and accurate classification results with minimum error. On the other hand, boosting being an ensemble classifier ensures accurate and precise results by integrating or combining set of weak classifier to form strong classifier. With the strong classifier outcomes final prediction results as normal or abnormal condition with minimal error rate is said to be achieved. Therefore, the ensembles of a set of classifiers are illustrated in Figure 3.

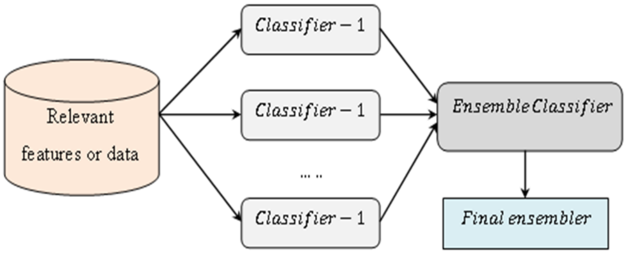

Fig 3. Structure of ensembles of a set of classifiers

Figure 3 given above shows the structural design of the Quadratic Weighted Entropy Boosting Classification model to enhance the classification performance of a set of training samples (i.e., using features selected '$FS = FS_1,\ FS_2,\ FS_3,\ \ldots, FS_n$' from the cardiovascular disease dataset 'DS'). The Quadratic Weighted Entropy Boosting Classification uses linear weighting mechanism to obtain strong classifier. The strong classifier is elucidated as given below.

$$f(FS) = \sum_{t=1}^{T} \delta_t BC_t(FS) \tag{6}$$

$$SC(FS) = Sign\left[f(FS)\right] \tag{7}$$

From the above Equations (6) and (7), ‘BC_t’ denotes the basis classifier, ‘δ_t’ the coefficient and ‘SC’ representing the strong classifier with the aid of relevant features selected ‘FS’ respectively. Then, given the features selected ‘FS’ and class labels ‘$((FS_1, y_1), (FS_2, y_2), (FS_3, y_3), \ldots (FS_n, y_n)]$’, strong classifier is mathematically formulated by selecting weak classifier with the smallest weighted error.

$$BC_t = argmin\,(\varepsilon_t) = \sum_{i=1}^{n} W_t\,(i)\,[y_i \neq BC_j\,(FS_i)] \tag{8}$$

From the above Equation (8), base classifier ‘BC_t’ with the minimum error ‘$argmin\,(\varepsilon_t)$’ is formulated by utilizing the initialized weight ‘$W_t\,(i)$’ with respect to the features selected ‘FS_i’. Then maximum likelihood ratio using quadratic classifier between testing and training samples (i.e., features selected) are utilized to obtain classification outcomes. The weak classifier has significant amount of training errors during classification. With the purpose of reducing the error rate, the ensemble model combines the set of weak classification outcomes. As there exists only two classes ‘$Normal$’ or ‘$Abnormal$’, with means ‘$\mu_0,\ \mu_1$’ and variances ‘σ_0, σ_1’, corresponding to ‘$y < 1\ [normal]\ and\ y > 1\ [abnormal]$’ respectively. Then, the quadratic maximum likelihood ratio is mathematically stated as given below.

$$LR = \frac{\sqrt{2\Pi\,|\sigma_1|}exp\left[-\frac{1}{2}\,(FS-\mu_1)^T\,\sigma_1^{-1}\,(FS-\mu_1)\right]}{\sqrt{2\Pi\,|\sigma_0|}exp\left[-\frac{1}{2}\,(FS-\mu_0)^T\,\sigma_0^{-1}\,(FS-\mu_0)\right]} \tag{9}$$

From the above Equation (9), the likelihood ratio ‘LR’ is estimated based on the two means ‘μ_1’, ‘μ_0’ and two variances ‘σ_1’, ‘σ_0’ for the corresponding features selected ‘FS’. Followed by the likelihood ratio, the weights are updated as given below.

$$W_{t+1}\,(i) = \frac{W_t\,(i)\,exp\,[-\delta_t y_i BC_t\,(FS_i)]}{\varepsilon_t} \tag{10}$$

From the above Equation (10), the updated weight ‘$W_{t+1}\,(i)$’, is arrived at based on the prior weight ‘$W_t\,(i)$’ and the assigned acceptable error rate ‘ε_t’ for the respective feature selected based classifier ‘$BC_t\,(FS_i)$’. Finally, the generalization error is estimated as the squared difference between actual and predicted results as given below,

$$Err = [y_a - y_p]^2 \tag{11}$$

From the above Equation (11), the generalization error ‘Err’ is evaluated by employing the actual output ‘y_a’ and the predicted classified output ‘y_p’ respectively. The pseudo code representation of Quadratic Weighted Entropy Boosting Classification is given below.

Algorithm 2. Quadratic Weighted Entropy Boosting Classification

```
Input: Dataset ‘$DS$’, Feature ‘$F = F_1, F_2, \ldots, F_n$’, Feature type ‘$T = 1, 2, 3$’, Feature type 1 ‘$T_1 = O$’, Feature type 2 ‘$T_3 = E$’, Feature type 3 ‘$T_3 = S$’
Output: Error minimized feature selection-based disease prediction
Step 1: Initialize features selected ‘$FS$’, Weight ‘$W$’, coefficient ‘$\delta_t$’
Step 2: Begin
Step 3: For each Dataset ‘$DS$’ with features selected ‘$FS$’
Step 4: Mathematically define strong classifier as given in Equations (6) and (7)
Step 5: Model base classifier ‘$BC_t$’ with the minimum error as in Equation (8)
Step 6: Evaluate quadratic likelihood ratio as in Equation (9)
Step 7: Update weight as in Equation (10)
Step 8: Evaluate generalization error as in Equation (11)
Step 9: If ‘$exp\,[-\delta_t y_i BC_t\,(FS_i)] \leq 1$’
Step 10: Then ‘$y_i = normal$’
Step 11: End if
Step 12: If ‘$exp\,[-\delta_t y_i BC_t\,(FS_i)] > 1$’
Step 13: Then ‘$y_i = abnormal$’
Step 14: End if
Step 15: End for
Step 16: End
```

As given in the above Quadratic Weighted Entropy Boosting Classification algorithm, for each training data (i.e., with the aid of features selected), ensemble model constructs a number of weak learners. The weak learner measures the likelihood ratio by means of quadratic classified between the testing and training samples. Upon higher likelihood the patient data is categorized into a heart disease class (i.e., abnormal) and vice versa. Then the ensemble model combines weak learners by assigning weight for each weak learner. Next, the generalization error is estimated to identify the classification results with minimum error. Owing to this result, the patient data are correctly predicted with minimum error rate.

3 Results and Discussion

In this section, the experimental evaluation of the proposed Reqularized Principle Component Quadratic weighted Entropy Boosting is compared with the existing benchmark methods such as Swarm Artificial Newral Networt (Swarm-ANN)[1] and Heart Disease Prediction Framework[2]. RPC-QWEB method is implemented in Python high level general purpose programming language. In order to perform the experimental setup, the cardiovascular disease dataset is applied for accurate heart disease diagnosis.

The dataset consists of three types of input features with 70000 patient records. The main objective of employing this dataset remains in identifying the outcomes pertaining to patients with heart disease. The dataset contains attributes such as age, height, weight, gender, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoking, alcohol intake, physical activity, and presence or absence of cardiovascular disease.

Among multiple features, relevant features (i.e., 10 features) are selected by means of Regularized Principal Component Regressive Feature Selection algorithm. Followed by which classification between normal and abnormality is estimated by utilizing Quadratic Weighted Entropy Boosting Classification algorithm. Finally, experimental evaluation is carried out for factors such as heart disease prediction accuracy, heart disease prediction time, sensitivity and error rate with respect to distinct numbers of patient data.

3.1 Comparison between different methods based on heart disease prediction accuracy

In this section, the performance evaluation of heart disease prediction accuracy is made. The heart disease prediction accuracy refers to the accuracy arrived at while performing prediction. The heart disease prediction accuracy is mathematically stated as given below.

$$HDP_{acc} = \frac{P_{AP}}{P_D} * 100 \tag{12}$$

From the above Equation (12), heart disease prediction accuracy 'HDP_{acc}' is measured on the basis of the patients' data involved in the simulation 'P_D' and the patients sample data accurately predicted 'P_{AP}'. It is measured in terms of percentage. Table 2 given below shows the heart disease prediction accuracy obtained for three different methods, RPC-QWEB, existing Heart Disease Prediction Framework (HDPF)[1], Swarm Artificial Neural Network (Swarm-ANN)[2] respectively.

Table 2. Heart disease prediction accuracy of different classifiers with cardiovascular disease dataset

Patients	Heart Disease Prediction Accuracy (%)		
	RPC-QWEB	Swarm-ANN	HDPF
500	97.42	95.91	93.87
1000	97.05	94.25	92.35
1500	96.85	93.85	91.55
2000	96.45	93.55	91.35
2500	96.25	93.15	91
3000	96	93	90.35
3500	95.85	92.55	90
4000	95.35	92.15	89.25
4500	95.15	92	89
5000	95	91.35	88.15

Figure 4 given below shows the graphical representation of heart disease prediction accuracy with respect to 5000 different patient's data. From the above figure, the accuracy is found to be improved using RPC-QWEB upon comparison with HDPF[1]

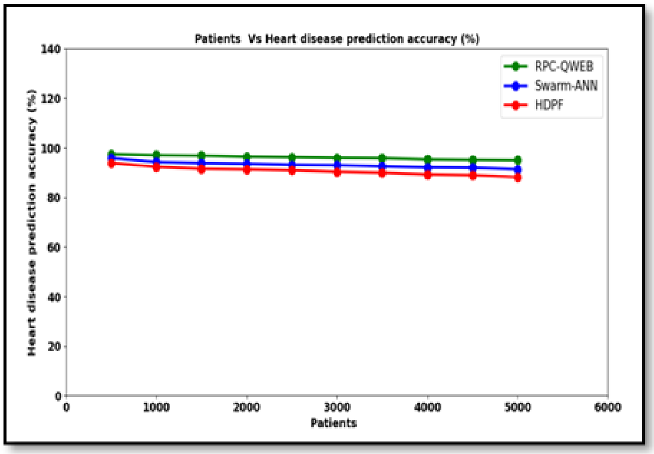

Fig 4. Comparisons in terms of heart disease prediction accuracy

and Swarm-ANN(2) respectively. The significant improvement in accuracy is due to the selection of relevant and significant features by employing Regularized Principal Component Regressive Feature Selection algorithm. By applying this algorithm, latent features were first obtained and then linear regression coefficients were employed for further processing. Finally, eigen value-one criterion were measured via conditionality checking. With this significant features were selected forming the basis for further classification. With this the heart disease prediction accuracy using RPC-QWEB is said to be improved by 3% compared to Swarm-ANN(1) and 6% compared to HDPF(2).

3.2 Comparison between different methods based on heart disease prediction time

Second, in this section, heart disease prediction time is measured. This is owing to the reason that while performing the prediction a significant amount of time is said to be involved. The heart disease prediction time is mathematically stated as given below.

$$HDP_{time} = \sum_{i=1}^{n} P_i * Time\ [FS] \tag{13}$$

From the above Equation (13), the heart disease prediction time 'HDP_{time}' is measured on the basis of the sample patient data considered 'P_i' and the time involved in the overall feature selection '$Time\ [FS]$' process. It is measured in terms of milliseconds (ms). Table 3 given below provides the heart disease prediction time results obtained from Equation (13) for all the three methods, RPC-QWEB, Swarm-ANN(1) and HDPF(2).

Table 3. Heart disease prediction time of different classifiers with cardiovascular disease dataset

Patients	Heart Disease Prediction Time (ms)		
	RPC-QWEB	Swarm-ANN	HDPF
500	415.5	472.5	492.5
1000	485.25	510.45	555.15
1500	515.35	545.55	590.45
2000	585.15	625.35	685.35
2500	625.45	685.15	745.45
3000	685.35	735.35	795.35
3500	725.45	780.45	845.15
4000	755.95	815.35	900.35
4500	825.15	845.15	945.15
5000	850.45	900.35	990.35

Figure 5 shows the graphical representation of heart disease prediction time using the proposed RPC-QWEB method and two state-of-the-art methods, Swarm-ANN(1) and HDPF(2). A steep increase in prediction time is found using all the three

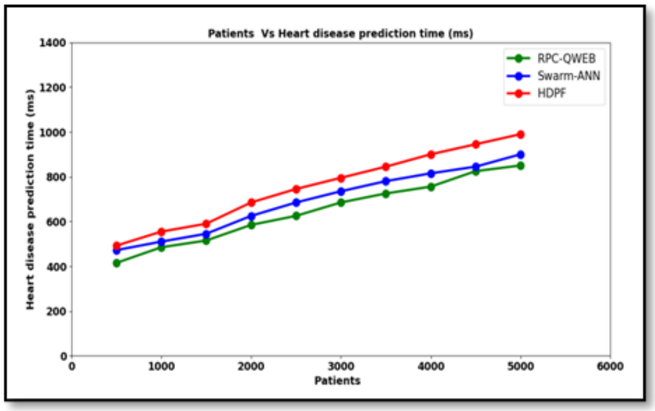

Fig 5. Comparisons in terms of heart disease prediction time

methods, though minimum time consumption is said to be observed using proposed RPC-QWEB method. The prediction time of heart disease using proposed RPC-QWEB method is significant reduced comparatively than[1] and[2]. This is owing to the application of geometric projection that in turn maps efficiently between three types of input features and overall features so that relevant features are selected in an extensive manner. As a result, the heart disease prediction time using RPC-QWEB method is said to be reduced by 7% compared to[1] and 14% compared to[2].

3.3 Comparison between different methods based on sensitivity

Third, in this section, the sensitivity involved in heart disease prediction is measured. Sensitivity refers to the accuracy of heart disease prediction test that reports the presence or absence of condition (i.e., heart disease). In other words, sensitivity is a measure of how well a test can identify true positives.

$$Sen = \frac{TP}{TP+FN} \tag{14}$$

From the above Equation (14), sensitivity rate '*Sen*', is estimated based on the number of true positives 'TP' and number of false negatives 'FN' respectively. Table 4 given below shows the sensitivity rate using three different methods, RPC-QWEB, Swarm-ANN[1] and HDPF[2].

Table 4. Sensitivity rate of different classifiers with cardiovascular disease dataset

Patients	Sensitivity		
	RPC-QWEB	Swarm-ANN	HDPF
500	0.97	0.94	0.92
1000	0.95	0.91	0.87
1500	0.93	0.89	0.85
2000	0.92	0.88	0.83
2500	0.91	0.87	0.81
3000	0.9	0.86	0.80
3500	0.88	0.84	0.79
4000	0.87	0.83	0.78
4500	0.85	0.81	0.77
5000	0.84	0.80	0.77

Figure 6 given below shows the sensitivity rate for three heart disease prediction methods, RPC-QWEB, Swarm-ANN[1] and HDPF[2]. As the objective of the proposed RPC-QWEB method is to identify everyone who has heart disease or to predict person with heart disease correctly, that number of false negatives should be less that necessitate high sensitivity. In other words to be more specific that is people suffering from heart disease should be highly likely to be identified as such by the test. From

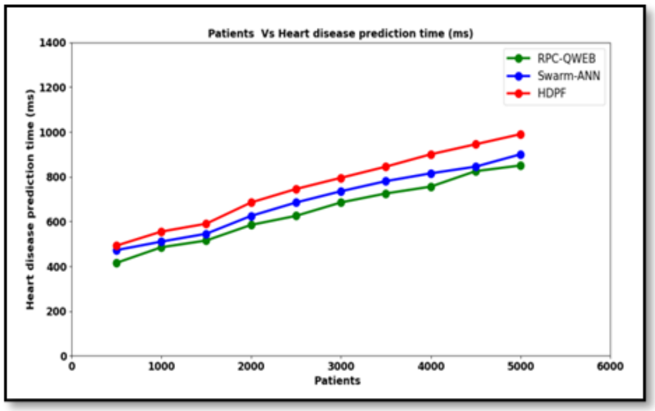

Fig 6. Performance comparisons in terms of sensitivity

the figure, the sensitivity rate is said to be comparatively higher using RPC-QWEB upon comparison with Swarm-ANN(1) and HDPF(2). This is owing to the lower false negative rate when applied with RPC-QWEB method than(1) and(2). The lower false negative rate using RPC-QWEB method was due to the application of Weighted Entropy Boosting that in turn combined the results of the weak classifiers to obtain strong classification results. As a result, the sensitivity rate using RPC-QWEB method is said to be improved by 5% compared to(1) and 10% compared to(2) respectively.

3.4 Comparison between different methods based on error rate

Finally, error rate is evaluated and provided as results in this section. During the prediction of heart disease a small amount of error is said to occur. This error rate is mathematically stated as given below.

$$ER = \sum_{i=1}^{n} \frac{P_{WD}}{P_i} * 100 \tag{15}$$

From the above Equation (15), the error rate 'ER', is measured on the basis of the patient's sample data 'P_i' involved in the simulation process and the patient's data wrongly detected 'P_{WD}'. It is measured in terms of percentage (%). Table 5 given below lists the error rate obtained by substituting the values in Equation (15) for three different methods, RPC-QWEB, Swarm-ANN(1) and HDPF(2).

Table 5. Error rate of different classifiers with cardiovascular disease dataset

Patients	Error Rate (%)		
	RPC-QWEB	Swarm-ANN	HDPF
500	3	4.8	6.6
1000	3.25	5.15	6.85
1500	3.85	5.3	7
2000	4	5.45	7.15
2500	4.15	5.75	7.35
3000	4.25	5.95	7.55
3500	4.4	6.15	7.85
4000	4.85	6.35	8.15
4500	5	6.55	8.3
5000	5.15	6.85	8.55

Finally, Figure 7 given below illustrates the graphical representation of error rate for 5000 sample patient data using proposed RPC-QWEB and existing state-of-the-art methods, Swarm-ANN(1) and HDPF(2) respectively. From the above figure, increasing the patients sample data causes an increase in the features for analysis and therefore a significant error rate is also observed. However, comparative analysis with(1) and(2) showed betterment using RPC-QWEB method. The reason behind the

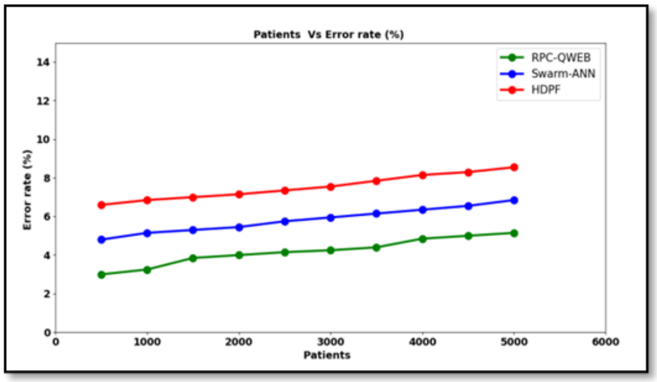

Fig 7. Performance comparisons in terms of Error rate

improvement was owing to the application of Quadratic Weighted Entropy Boosting Classification algorithm. By applying this algorithm, the generalization error was evaluated as the squared difference between actual and predicted results. Also, with the utilization of quadratic maximum likelihood ratio, for the respective features, the weights were also updated. As a result, the error rate using RPC-QWEB method was said to be reduced by 29% compared to [1] and 45% compared to [2] respectively.

Heart disease prediction accuracy is improved by using Regularized Principal Component Regressive Feature Selection algorithm and Sensitivity rate is enhanced by the application of Weighted Entropy Boosting that combined the results of weak classifiers for attaining strong classification results. Generalization error was evaluated as squared difference between actual and predicted results for reducing the error rate.

4 Conclusion

In this paper, a Regularized Principal Component and Quadratic Weighted Entropy Boosting (RPC-QWEB) for predicting heart disease is designed based on the Regularized Principal Component and Quadratic Weighted Entropy Boosting (RPC-QWEB) method. The main purpose of this work is to receive a large set of monitored parameters from cardiovascular disease dataset and predicts disease using the proposed RPC-QWEB method. In the initial phase, the proposed RPC-QWEB method obtains three types of input features from the raw cardiovascular disease dataset for training and evaluating the dataset. In the next phase, computationally efficient robust and accurate features are selected by means of Regularized Principal Component Regressive Feature Selection algorithm. Here, linear regression and geometric progressions are applied to the input features by evaluating the eigen value to select relevant features. Finally, Quadratic Weighted Entropy Boosting Classification algorithm is applied to the selected features by employing quadratic likelihood ratio as the weak classifier and combining their results by means of modifying the weight via generalization error. The accuracy, time, sensitivity and error rate of the proposed method is validated by a benchmark heart disease dataset with twelve features. This approach results in a 7% reduction in prediction time, a 5% increase in sensitivity, a 26% drop in error rate, and a 3% improvement in prediction accuracy when compared to the Swarm-ANN method. Likewise, our approach yields a 6% gain in prediction accuracy, a 10% increase in sensitivity, a 14% reduction in prediction time, and a 45% reduction in error rate when compared to the HDPF method. The simulation results exhibit that the proposed RPC-QWEB method outperforms the conventional learning methods in terms of numerous performance matrices. The future direction of the work is to perform an efficient heart disease prediction by using deep learning and machine learning methods with higher accuracy and lesser time consumption. We may also consider some other performance metrics such as False positive rate, F1 Score and so on.

References

1) Nandy S, Adhikari M, Balasubramanian V, Menon VG, Li X, Zakarya M. An intelligent heart disease prediction system based on swarm-artificial neural network. *Neural Computing and Applications*. 2023;35(20):14723–14737. Available from: https://doi.org/10.1007/s00521-021-06124-1.
2) Ashri SEA, El-Gayar MM, El-Daydamony EM. HDPF: Heart Disease Prediction Framework Based on Hybrid Classifiers and Genetic Algorithm. *IEEE Access*. 2021;9:146797–146809. Available from: https://doi.org/10.1109/ACCESS.2021.3122789.
3) Mohan SK, Thirumalai C, Srivastava G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *IEEE Access*. 2019;7:81542–81554. Available from: https://doi.org/10.1109/ACCESS.2019.2923707.

4) An Y, Huang N, Chen X, Wu F, Wang J. High-Risk Prediction of Cardiovascular Diseases via Attention-Based Deep Neural Networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2021;18(3):1093–1105. Available from: https://doi.org/10.1109/TCBB.2019.2935059.

5) Dutta A, Batabyal T, Basu M, Acton ST. An efficient convolutional neural network for coronary heart disease prediction. *Expert Systems with Applications*. 2020;159:113408. Available from: https://doi.org/10.1016/j.eswa.2020.113408.

6) Wong ND, Zhao Y, Xiang P, Coll B, López JAG. Five-Year Residual Atherosclerotic Cardiovascular Disease Risk Prediction Model for Statin Treated Patients With Known Cardiovascular Disease. *The American Journal of Cardiology*. 2020;137:7–11. Available from: https://doi.org/10.1016/j.amjcard.2020.09.043.

7) Zhou C, Li A, Hou A, Zhang Z, Zhang Z, Dai P, et al. Modeling methodology for early warning of chronic heart failure based on real medical big data. *Expert Systems with Applications*. 2020;151:113361. Available from: https://doi.org/10.1016/j.eswa.2020.113361.

8) Shankar V, Kumar V, Devagade U, Karanth V, Rohitaksha K. Heart Disease Prediction Using CNN Algorithm. *SN Computer Science*. 2020;1(3). Available from: https://doi.org/10.1007/s42979-020-0097-6.

9) Kumar Y, Koul A, Singla R, Ijaz MF. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *Journal of Ambient Intelligence and Humanized Computing*. 2023;14(7):8459–8486. Available from: https://doi.org/10.1007/s12652-021-03612-z.

10) Haithamelwahsh E, El-Shafeiy S, Tawfeek MA. A new smart healthcare framework for real-time heart disease detection based on deep and machine learning". *Peer J Computer Science*. 2021;7:1–34. Available from: https://doi.org/10.7717/peerj-cs.646.

11) Patro SP, Nayak GS, Padhy N. Heart disease prediction by using novel optimization algorithm: A supervised learning prospective. *Informatics in Medicine Unlocked*. 2021;26:1–17. Available from: https://doi.org/10.1016/j.imu.2021.100696.

12) Mehmood A, Iqbal M, Mehmood Z, Irtaza A, Nawaz M, Nazir T, et al. Prediction of Heart Disease Using Deep Convolutional Neural Networks". *Arabian Journal for Science and Engineering*. 2021;46:3409–3422. Available from: https://doi.org/10.1007/s13369-020-05105-1.

13) Bertsimas D, Mingardi L, Stellato B. Machine Learning for Real-Time Heart Disease Prediction. *IEEE Journal of Biomedical and Health Informatics*. 2021;25(9):3627–3637. Available from: https://doi.org/10.1109/JBHI.2021.3066347.

14) Yazdani A, Varathan KD, Chiam YK, Malik AW, Ahmad WAW. A novel approach for heart disease prediction using strength scores with significant predictors. *BMC Medical Informatics and Decision Making*. 2021;21(1):1–16. Available from: https://doi.org/10.1186/s12911-021-01527-5.

15) Garate-Escamila AK, Hassani AHE, Andres E. Classification models for heart disease prediction using feature selection and PCA. *Informatics in Medicine Unlocked*. 2020;19:1–11. Available from: https://doi.org/10.1016/j.imu.2020.100330.

16) Sushma SJ, Assegie TA, Vinutha DC, Padmashree S. An improved feature selection approach for chronic heart disease detection. *Bulletin of Electrical Engineering and Informatics*. 2021;10(6):3501–3506. Available from: https://doi.org/10.11591/eei.v10i6.3001.

17) Krishnamoorthi R, Joshi S, Almarzouki HZ, Shukla PK, Rizwan A, Kalpana C, et al. A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques. *Journal of Healthcare Engineering*. 2022;2022:1–10. Available from: https://doi.org/10.1155/2022/1684017.

18) Chen RC, Dewi C, Huang SW, Caraka RE. Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*. 2020;7:1–26. Available from: https://doi.org/10.1186/s40537-020-00327-4.

19) Yahaya L, Oye ND, Garba EJ. A Comprehensive Review on Heart Disease Prediction Using Data Mining and Machine Learning Techniques". *American Journal of Artificial Intelligence*. 2020;4(1):20–29. Available from: https://doi.org/10.11648/j.ajai.20200401.12.

20) Zhang D, Chen Y, Chen Y, Ye S, Cai W, Jiang J, et al. Heart Disease Prediction Based on the Embedded Feature Selection Method and Deep Neural Network. *Journal of Healthcare Engineering*. 2021;2021:1–9. Available from: https://doi.org/10.1155/2021/6260022.

21) Ali F, El-Sappagh S, Islam SMR, Kwak D, Ali A, Imran M, et al. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Information Fusion*. 2020;63:208–222. Available from: https://doi.org/10.1016/j.inffus.2020.06.008.

22) Mamun MMRK, Elfouly T. Detection of Cardiovascular Disease from Clinical Parameters Using a One-Dimensional Convolutional Neural Network. *Bioengineering*. 2023;10(7):1–29. Available from: https://doi.org/10.3390/bioengineering10070796.