Journal of Computer Science

Original Research Paper

An Analysis of Heart Disease for Diabetic Patients Using Recursive Feature Elimination with Random Forest

Bindushree Doddasiddavanahalli Channabasavaraju and Udayarani Vinayakamurthy

School of Computing and Information Technology, REVA University, Bangalore, India

Article history
Received: 30-09-2019
Revised: 20-11-2019
Accepted: 24-01-2020

Corresponding Author:
Bindushree
Doddasiddavanahalli
Channabasavaraju
School of Computing and
Information Technology,
REVA University, Bangalore,
India
Email: binduojin@gmail.com

Abstract: Data Mining Techniques are used in many areas like banking, healthcare, education because of extracting relevant information from the database. Many algorithms in data mining are used to predict the disease to reduce the patient treatment cost as well as increase the diagnosis accuracy. In this research work, the risk level of diabetic heart disease patient can be predicted using two datasets such as Pima Indians Diabetes dataset and heart disease dataset. Based on the World Health Organization (WHO), diabetes is the one of the biggest health concerns so mining the diabetes data is an ambiguous task. Diabetic patients may also have to suffer from other diseases like heart disease, eye complications, kidney disease, nerve damage, foot problems, skin complications and dental diseases. However, the existing techniques faced the difficulty to identify inconsistent and redundant features. In this research work, the Random Forest Feature Selection Technique (RFS) is developed to identify the significant features and eliminate irrelevant features which are used to improve the predicting accuracy of Cardio Vascular (CV) disease for diabetic patients. According to the output from the extensive experiments, this research categorizes that whether the diabetic patients having CV or not by using True Positive as diabetes and True Negative as no diabetes.

Keywords: Data Mining, Disease Prediction, Diabetes, Feature Selection, Heart Disease, Random Forest

Introduction

Data mining is used to analyze a number of datasets and extracts data with classification techniques. It is used to predict the patterns and trends for decision making (Zhang *et al*., 2017; Amin *et al*., 2019). The major aim of this research work is to predict whether diabetic patients will get chance to heart disease or not. The efficiency of data mining can be affected because of the selection of features and the techniques used to predict the disease among patients. In addition to this, this technique faces the difficulty to identify inconsistent and redundant features without appropriate preparations that are present in medical datasets of healthcare industry (Maji and Arora, 2019; Looker *et al*., 2015). The presence of inconsistency and data redundancy in raw data provides the poor performance of the traditional algorithm, which is stated in the existing work (Kavitha and Kannan, 2016). Therefore, an effective preprocessing techniques are introduced to remove these redundant data. Even though the performance of machine learning techniques is increased, the unwanted features in the dataset affects their efficiency (Long *et al*., 2015). To achieve high accuracy in predicting heart disease, a proper feature selection method is used along with data preparation by using significant features (Hidalgo *et al*., 2017).

Researchers face the problem of combining the proper set of features with suitable data mining technique, even though it is identified that the feature selection plays an important role in selecting the appropriate technique. Moreover, the studies present less attention towards the prediction of risk factors of CV diseases for diabetic patients by using significant features (Sharma and Saxena, 2017; Prakash *et al*., 2018; Kang *et al*., 2015). The performance of the prediction models is decreased due to the difficulty of identifying the suitable combination of significant features. The achievement for predicting the CV disease with high accuracy is not an easy task, which is stated by existing work (Shouman *et al*., 2013). But, definitely the selection of significant features will improve the prediction accuracy, which proved that

Science Publications

the necessary of identifying the significant features is important to achieve the goal (Guo and Garvey, 2015; Cui *et al*., 2018; Bindushree, 2016; Bindushree and Rani, 2017). The difficulty to identify inconsistent and redundant features was the major drawback of the existing methods. The major aim of this research work is to select the risk factors for heart problem prediction in diabetic patients by using RFS. The research work uses the normalization process to pre-process the patient data for diabetes and heart problem prediction. The importance of features can be selected by removing the irrelevant data using Recursive Feature Elimination (RFE) technique to improve the higher classification accuracy. The RF classification algorithm is designed to diagnosis whether the diabetic patients will get a chance to heart disease. The objective of the proposed Random Forest Feature Selection Technique (RFS) is to identify the significant features which are used to improve the predicting accuracy of Cardio Vascular (CV) disease for diabetic patients on the basis of the probability of disease severity. The main advantage of the proposed system is that it considers all the important functions for the predicting the heart disease risk factors for diabetic patients and subsequently it is greater suitable for making recommendations based totally on deductive inference. The experimental results are conducted on two standard datasets from UCI, which is used to validate the efficiency of RFS-RFE algorithm when compared with other existing methods.

The rest of the paper is arranged as follows: Section 2 describes the survey of recent year techniques for predicting heart disease or diabetes for a patient. The proposed methodology of this research work is explained in section 3. The experimental results that validate the efficiency of proposed RFS-RFE are represented in section 4. The conclusion with future work is given in section 5.

Literature Review

Numerous methodologies are developed by researchers in predicting diabetic heart diseases. In this section, a survey of recent techniques is discussed which is used to predict the CV and diabetes diseases among the patients.

Vivekanandan and Iyengar (2017), performed the feature selection and optimized the selected features by using modified Differential Evolution (DE) algorithm for CV disease. A Fuzzy Analytic Hierarchy Process (FAHP) and Feed-Forward Neural Network (FFNN) were used to predict the heart disease with selected critical features. The validation of DE, FFNN and FAHP were carried out by extensive experiments on publicly available dataset using parameter metrics such as sensitivity, accuracy and specificity. In DE method, there was a greater deviation in average execution time because of high error values which were used for predicting the heart disease.

Chen *et al*. (2019), analyzed the asymptomatic diabetic patients to identify the relationship between CV, bilirubin and All-Cause Mortality (ACM). The actual and serum bilirubin values were negatively correlated with CV death and ACM, but higher serum of bilirubin values was associated with less risk for CV and ACM in diabetic patients. The risk factor for predicting the CV was high when compared with ACM people due to improvement in bilirubin. Although diabetic patients measured the concentrations of fasting and postprandial glucose, they didn't measure the HbA1C routinely. But, there was a linear and predictable relationship between glucose and HbA1C to predict the disease accurately.

Bhuvaneswari and Manikandan (2018), implemented the hybrid techniques namely Temporal Fuzzy Ant Miner Tree (TFAMT) and temporal feature selection for making an effective decision on the analysis of type-2 diabetes. The temporal weighted genetic algorithm was used to improve the detection accuracy by preprocessing the image and text data. In the extracted regulations, the variety of functions were reduced by fuzzy rule extractor. The TFAMT considered all the important functions for diabetic analysis and make a suitable recommendation which depends on totally deductive inference information found out by fuzzy rules. However, the presence of additional features which made confusions in decision making which lead to increase the selection time and the reduction in accuracy.

Mdhaffar *et al*. (2017), designed a Complex Event Processing (CEP) technology with statistical approaches for analyzing the heart failure prediction in patients. The threshold customization approach was used to calculate the threshold values automatically and updated at runtime, which was the novelty of CEP. The efficiency of CEP technology was evaluated by experimental results in terms of precision and recall. The time for processing and deploying the rules was quite low which was illustrated by results of CEP. Based on real-world cases, the CEP method provides poor performance on large-scale data for identifying heart failure.

Novara *et al*. (2016), the input signals such as emotions, food and physical activity were effectively recovered by blind identification approach for type 1 diabetic patients. The samples for five type 1 diabetic patients were collected for experimental study to the blood glucose for that five patients. The populations of patients were effectively deal with this approach to characterize the significant variability in terms of life habits and metabolic processes. In the simulation test, the method verified the glucose concentration was very large for all patients when no insulin was injected.

From the above survey, the existing methods focused only on either heart disease or diabetic prediction, but the proposed method tried to predict that whether the diabetic patients will have sudden cardiac arrest or not by using Random Forest classifier.

Proposed Methodology

In this section, the detail description of proposed RFE techniques to find the risk factors of diabetic patients for predicting the risks of heart disease is presented. Figure 1 shows the block diagram of proposed method. The major steps of the proposed method are pre-processing the raw data, split the data for training and testing, the important features are extracted by using random forest technique and the irrelevant data can be removed by using RFE technique. Also, the symptoms and diagnose of both heart and diabetic disease are discussed.

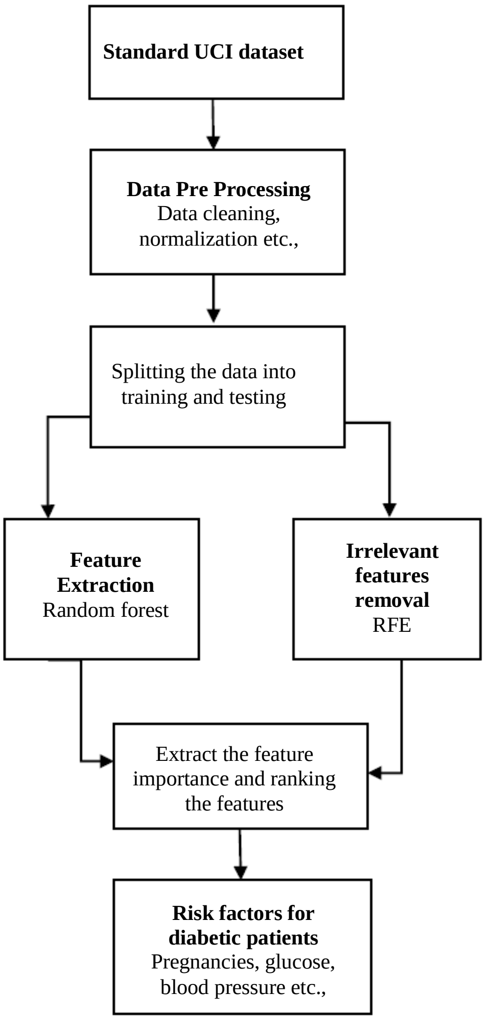
Standard UCI dataset
Data Pre Processing
Data cleaning, normalization etc.,
Splitting the data into training and testing
Feature Extraction
Random forest
Irrelevant features removal
RFE
Extract the feature importance and ranking the features
Risk factors for diabetic patients
Pregnancies, glucose, blood pressure etc.,

Fig. 1: Block diagram of proposed methodology

The raw data can be collected by using two standard UCI datasets for predicting the risk factors for diabetic patients whether they have a chance to get heart disease or not. After collecting the data, this research work develops a new dataset called diabetic-heart disease dataset by combining the common attributes in both datasets. These data may contain noises, outliers, missing values which leads poor classification accuracy. Hence, pre-processing should be done for this data to find the accurate risk factors for diabetic patients.

Pre-Processing

After data collection, the data is preprocessed because 6 records are having missing values, which should be removed from the dataset by reducing the number of records from 303 to 294. In this stage, the noises and missing values are reduced by several data pre-processing techniques like data cleaning, normalization, transformation, discretization and integration of data. In this research work, normalization method is used for pre-processing to reduce the dimensionality of data. In this process, the method normalized the variables into [0,1] intervals to avoid the effects of scale. When the data are extracted from the collected dataset, which is used for various process such as training, testing and validation. The normalization method employed for the dataset can be observed in the (Equation 1):

$$V_i = \frac{a_i - \min_{a_i}}{\max_{a_i} - \min_{a_i}} \tag{1}$$

where, a_i represents a value to normalize for the i^{th} variable, $\min_{a_i}$ is the minimum value registered for this variable in the training set and $\max_{a_i}$ is the maximum valor registered for this variable in the training set. After pre-processing stage, the data can be split into training and testing to obtain high risk factor. For experimental analysis, the pre-processed data can be split into 70% for training and 30% for testing.

Selection of Features

The most relevant features are selected in this process to study and reduce the original high dimensional feature vector. Feature selection is considered as one of the most important process for building an efficient and powerful prediction engine. The training data alone is utilized for the selection of features and testing data pretended to be unknown till the evaluation process is started. The importance of different feature components provides intuitive insights which makes the prediction process more transparent to its users. In this research work, the computational complexity is reduced by using these selected features, where RFS Technique is used to select the risk factors for predicting the heart disease in diabetic patients. According to the construction of random forest, the fusion of multiple decision trees is conducted and the values of random vector are obtained independently based on each tree's original feature space with same distribution. In decision tree, the different features are selected to form the tree nodes and their importance are calculated using random forest (Ritika and Mayank, 2016). In feature selection, still the presence of some unrelated data leads poor accuracy for prediction process. Therefore, the proposed RFE technique used to remove these irrelevant features from the selected data. The features remaining after a several iterations are deemed to be the most useful for discrimination. Therefore, the best features were iteratively dropped by using RFE instead of removing the worst features. In next section, the importance of features is discussed with RFE techniques to further predict the best features for diabetic patients.

Recursive Feature Elimination (RFE) Technique

The features are ranked using RFE selection method according to their important measures, which is considered as a recursive process. At each iteration feature importance are measured and the less relevant one is removed. A group of features is removed every time by using RFE method to accelerate the process. During the evaluation of various features subset (especially for highly correlated features) in stepwise elimination process, the recursion is important because the relative information for some measures are changed substantially. A final ranking are constructed by eliminating the features in the inverse order, where the first *n* features are selected only from this ranking by the feature selection process itself. Figure 2 shows the graphical representation of the importance of features which can be identified by using RFE.

The pseudo-code of RFE is shown below which is used to eliminate the irrelevant features and improves the risk factor of diabetic patients whether they will have heart attack or not.

Algorithm 1: Recursive Feature Elimination

```
Inputs:
        Training set T
        Set of $p$ features $F = \{f_1,\dots,f_p\}$
        Ranking method $M(T, F)$
Outputs:
        Final ranking $R$
Code:
        Repeat for $i$ in $\{1: p\}$
          Rank set using $F$, $M(T, F)$
                $f^* \leftarrow$ last ranked feature in $F$
                $R(p-i+1) \leftarrow f^*$
                    $F \leftarrow F - f^*$
```

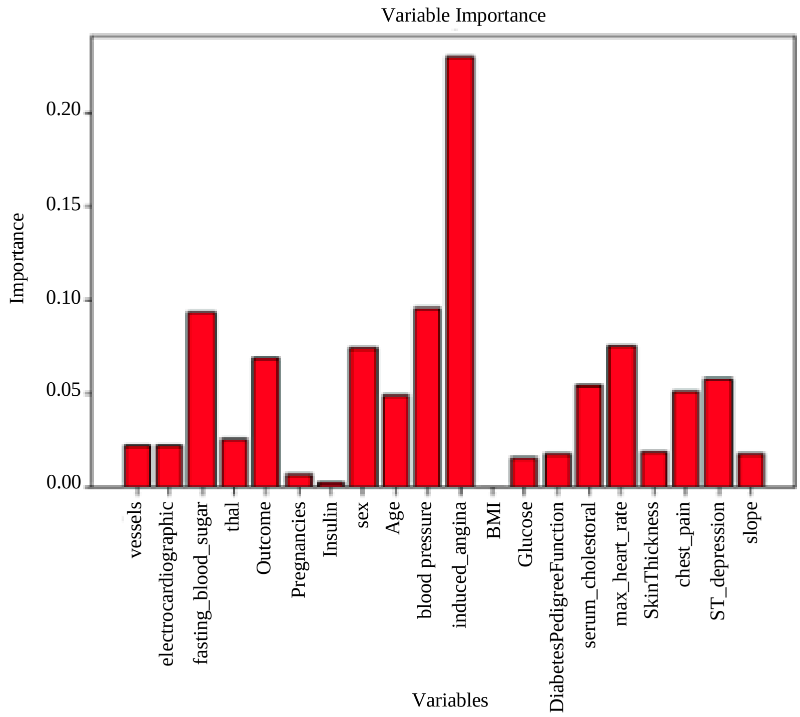

Fig. 2: Importance of selected features

Scaling the Factors

Next, the values of predicted attribute for the presence of heart disease in the dataset was transformed from multiclass values (0 for absence and 1, 2, 3, 4 for presence) to the binary values (0 for absence; 1 for presence of heart disease). The diagnosis values of 2 to 4 are converted into 1 by using the task of data pre-processing, where the resultant datasets contain only the diagnosis values as 0 and 1 for predicting the heart disease. The 134 records are assigned for “1” and 160 records are assigned for “0” from the total distribution of 294 records after the process of transformation and reduction. Then, the important features are ranked once scaling process is finished. Figure 3 shows the ranking importance of features which are used to predict whether the diabetic patients having heart disease or not.

By using the above cumulative graph, it is clearly explaining that more than five attributes gain the importance of ranking for further predicting. In this research work, the total number of 15 features gained the 95% feature importance. After gaining these important features, they can be given as an input for classifier techniques to predict the heart disease.

Classification Techniques

In this research work, three types of classifiers are used to identify the relation between heart disease and diabetic patients by using Support Vector Machine (SVM) and Random Forest (RF). SVM are now becoming standard tools in various areas such as healthcare to predict the diseases. According to the Structural Risk Minimization Principle, the SVM algorithm locates the hyper plane with the maximum margin, i.e., with the maximum distance from the hyper plane to the closest vector in each class. The multiclass problems with binary classifiers are solved by extending the SVM with various strategies for handling the multiclass datasets. The feature importance of RFE algorithm is considered by using the components of SVM which are presents in the implementation of multiclass SVM. Based on the risk factors alone, we are predicting the heart disease among the diabetic patients. The prediction includes classifying, the percentage of patients suffering from the heart disease. But, when comparing the efficiency of three algorithms, RF provides better performance in those selected features which gives higher classification accuracy are as follows.

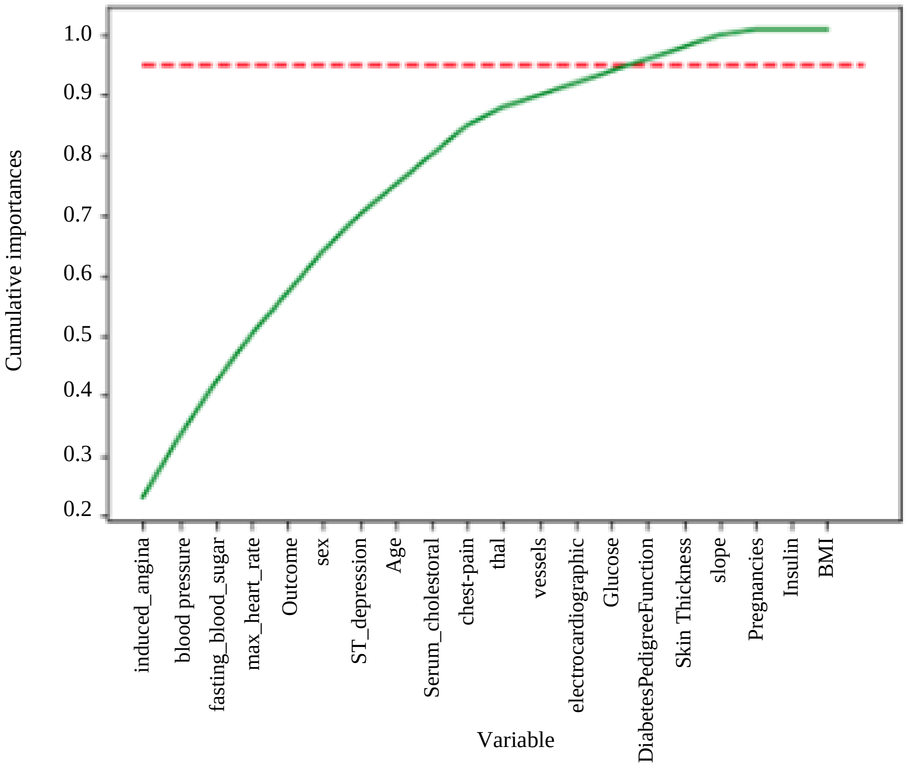

Fig. 3: Cumulative importance of features

Prediction using Radom Forest Technique

Among the research community, ensemble learning algorithms namely boosting, bagging and RF gained more interest due to its robustness against outliers and noises than single classifiers. In general, it uses hundreds of diverse classification trees as a composite classifier. While applying the majority rule of a sample over the votes of single classifiers, the decisions are carried out for selecting the final classification for that given sample. Each tree is grown by using a reduce samples to avoid the correlated and similar predictions in a training set. The best splits are identified by introducing the randomness in algorithm for increasing the diversity between the training samples. There are many advantages presents in the application of RF, which are developed by Breiman and their benefits are stated as below:

- It works on large scale dataset effectively and efficiently
- It can process more than thousands of input variables without deleting any variables in the tree
- It predicts which variables plays an important role in classification process
- It generates an internal unbiased estimate of the generalization error
- The pairs of cases are used to locate the outliers and their proximities are identified by RF algorithm
- It can able to handle the presence of noise and outliers
- When compared with other ensemble methods, RF is computationally light-weight method

A set of decision trees with controlled variations are constructed by combining the randomly selected features with “bagging” ideas of Breiman's.

Algorithm: Random forest classifier

Input: Consider the dataset N, which is a collection of observed and their associated class values in training process.

Output: Prediction of risk factors for heart disease among diabetic patients

According to the following steps, the construction of every tree are as follows:

1. Consider N as number of training cases and M as number of variables in the classifier
2. At a node of the tree, the decisions are identified by using the number of input variables as m and the variables in the classifier M should be higher than input variables m
3. From all N available training cases, times with replacement are chosen to select the training set for this tree. The errors of the trees are identified by using the remaining cases and also by calculating their classes
4. Randomly choose m variables for each node of the tree and according to these m variables, the best splits are calculated in the training set
5. To construct a normal tree classifier, each tree should not be pruned and new samples are pushed down the tree for prediction. It ended up when the training samples labels are assigned in the terminal node and the iterations are conducted over all trees in the ensemble. The RF predictions are reported by finding the average vote of all trees

While entering in the model, the relevance features are identified by RF in a way, where each features are shuffled at a time, then prediction error on this shuffled dataset are estimated by Out-Of-Bag (OOB). When alterations are conducted in this way, the prediction error of irrelevant features are not changing like relevant features. The shuffled features’ relevance is closely related to the relative loss in performance between the shuffled and original dataset. The feature importance measures are combined with RFE algorithm in RF-RFE technique. While OOB subsets are used by RF algorithm, the features importance is estimated and their computational efforts are not increased. In addition, a multiclass algorithm is developed by RF, which provides a better important measures than the combination of binary problems used in SVM. By using these RF classifier, this research work can predict whether the diabetic patients will be affected by sudden CV attack or not in future.

Experimental Analysis

In this section, the experimental analysis of RFE techniques is validated by using dataset created by this research work. The dataset can be created by combining the common attributes of two UCI datasets such as Pima Indians Diabetes dataset and heart disease dataset. After the preprocessing step, the training and testing set for the proposed model RFE is split into 70% for Training and 30% for Testing. The parameters such as accuracy, sensitivity, specificity and precision are used to validate the efficiency of proposed RFS-RFE techniques with Artificial Neural Network (Mathan *et al*., 2018) (ANN) with fuzzy and SVM. In this research work, the implementation of both ANN with fuzzy and SVM are carried out with RFE to predict whether the diabetic patients will have a heart attack or not.

Dataset Description

The proposed method tried to predict the sudden CV attack in future for diabetic patients by using two different datasets such as Heart disease prediction and Pima Indians Diabetes Database. The method chooses the common attributes from both datasets to predict whether the diabetic patient has heart attack or not. Table 1 and 2 shows the attributes of two datasets, whereas Table 3 explains the Diabetes-Heart Disease dataset created by proposed methodology.

Table 1: Diabetes dataset

Dataset	Total attributes	Number of instances
Pima indians diabetes dataset	Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, Outcome	768

Table 2: Heart disease dataset

Dataset	Total attributes	Number of instances
Heart disease dataset	Age, sex, chest-pain, blood pressure, serum-cholestoral, fasting-blood-sugar, electrocardiographic, max-heart-rate, induced-angina, ST-depression, slope, vessels, thal, diagnosis	303

Table 3: Diabetes-heart disease dataset

Dataset	Total attributes	Number of instances
Diabetes- heart disease dataset	BMI, Insulin, Pregnancies, Glucose, slope, Diabetes Pedigree Function, Skin Thickness, electrocardiographic, vessels, thal, Age, chest-pain, serum-cholestoral, ST-depression, Outcome	294

Attribute Selection

The common attributes for both diabetes and heart disease datasets such as age, sugar and pressure are extracted by using RFE, the new dataset will be created. The person must check the diabetic to predict the risk factors of heart disease and the outcomes may be yes or no. If yes, then the person is affected by the heart disease. If not, the result will depend on the output of diabetic dataset.

Parameter Evaluation

The presence of heart disease for diabetic patients are represented as "1", whereas the absence of disease is defined as "0" by using confusion matrix of proposed RFS-RFE model, that are explained in Table 4.

Accuracy

Prediction accuracy is defined as the percentage of test set tuples for the presence or absence of heart disease, which is correctly predicted. The Equation (2) describes the accuracy:

$$Accuracy = \frac{TP + TN}{Total\ No.of\ patients} \quad (2)$$

Precision

The proportion of true positive are defined among all positive tuples is known as precision and the mathematical expression of precision is given in Equation (3):

$$\Pr ecision = \frac{TP}{TP + FN} \quad (3)$$

Table 4: Confusion matrix

Output values	0	1
1	False Negative (FN)	True Positive (TP)
0	True Negative (TN)	False Positive (FP)

Where:

TP - Patient who really has heart disease and is diagnosed with heart disease
TN - Patient who does not have heart disease and is not diagnosed with heart disease
FN - Patient who really has heart disease but is diagnosed with the absence of heart disease
FP - Patient who does not have heart disease but is diagnosed with heart disease

Total Number of patients = TP + TN + FN + FP. Both FP and FN are incorrect classifications.

Sensitivity

Among all the positive tuples, the ratio of finding the correct positive tuples is defined as sensitivity or true positive rate, which is explained in Equation (4):

$$Sensivity = \frac{TP}{TP + FN} \quad (4)$$

Specificity

In the total number of negative tuples, the ratio of negative tuples that are identified correctly is defined as true negative rate or specificity, that is explained in Equation (5):

$$Specificity = \frac{TN}{TN + FN} \quad (5)$$

Performance of Proposed RFS-RFE

In this section, the performance of RFS are evaluated for three datasets such as Heart disease as HD, Diabetic disease dataset as DD and Diabetes-Heart Disease Dataset as DHD in terms of various parameters such as accuracy, precision, sensitivity and specificity. The proposed method is mainly focused to identify the risk factors of heart disease for diabetic patients. Therefore, the effectiveness of RFS is validated by selecting the features and without selecting the features in which the values are tabulated for various parameters. Table 5 validated the performance of RFS-RFE for accuracy values.

The Fig. 4 shows the graphical representation for accuracy values. It is clearly shows that the importance of features plays a major role for predicting the risk factors of heart disease. By selecting the features, the RFS-RFE achieved the accuracy of 83.49%, 78.57% and 82.56% for HD, DD and DHD datasets, whereas the accuracy of 80.2%, 75.03% and 80.69% for HD, DD and DHD by without selecting the important features. Table 6 and Fig. 5 presents the values for sensitivity of RFS-RFE for all the three datasets.

Here, the sensitivity is not affected by the importance of features to predict the risk factors. But, when compared with all other three datasets, the DD dataset provides poor performance when the features are not selected. When the important features are not selected, the RFS-RFE method achieved sensitivity of 70.21%, 42.85% and 65.45% for HD, DD and DHD. But, the same method achieved the sensitivity of 72.09%, 57.4% and 66.66% for all the three datasets by selecting the important features for predicting the risk factors. Table 7 and Fig. 6 presents the precision values of RFS-RFE for different datasets.

Table 5: Accuracy values of proposed RFS-RFE

	Accuracy	
Datasets	Without features	With features
HD	80.20	83.49
DD	75.03	78.57
DHD	80.69	82.56

Table 6: Sensitivity values of RFS-RFE

	Sensitivity	
Datasets	Without features	With features
HD	70.21	72.09
DD	42.85	57.40
DHD	65.45	66.66

Table 7: Precision of RFS-RFE

	Precision	
Datasets	Without features	With features
HD	83	84
DD	69	78
DHD	82	83

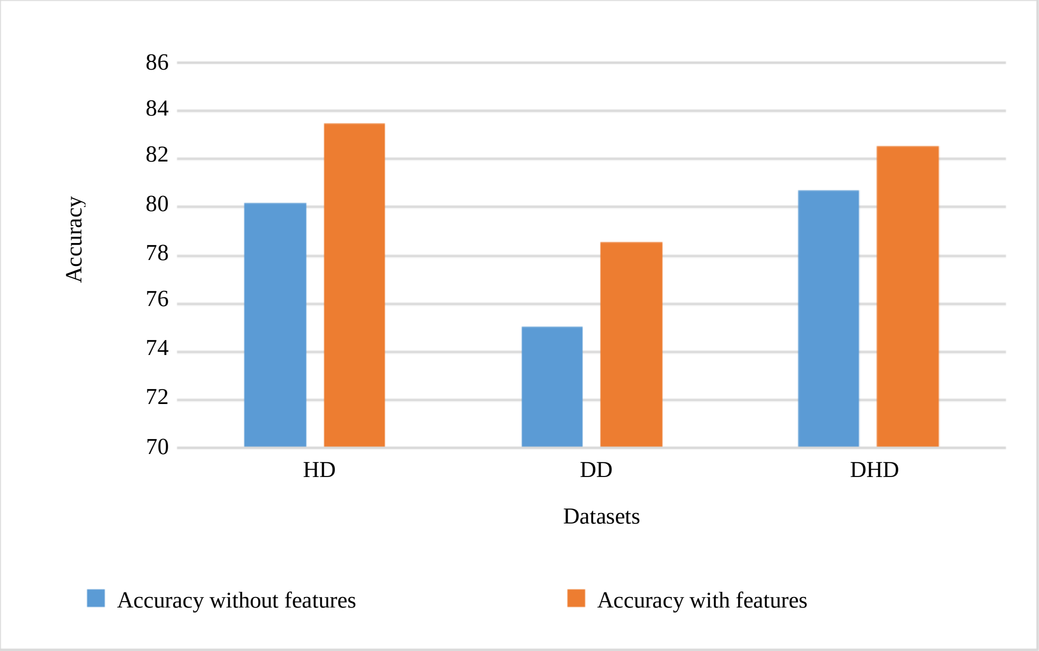

Fig. 4: Accuracy for proposed RFS-RFE

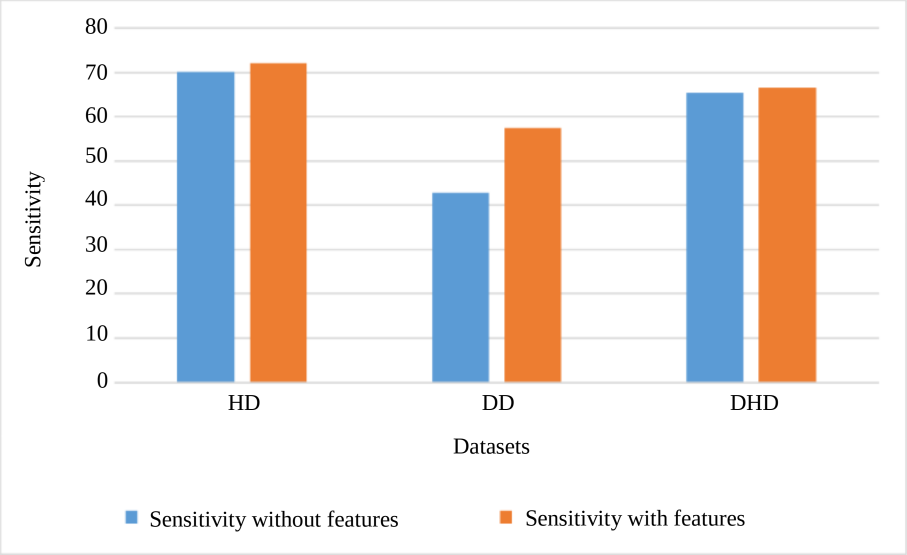

Fig. 5: Sensitivity of RFS-RFE

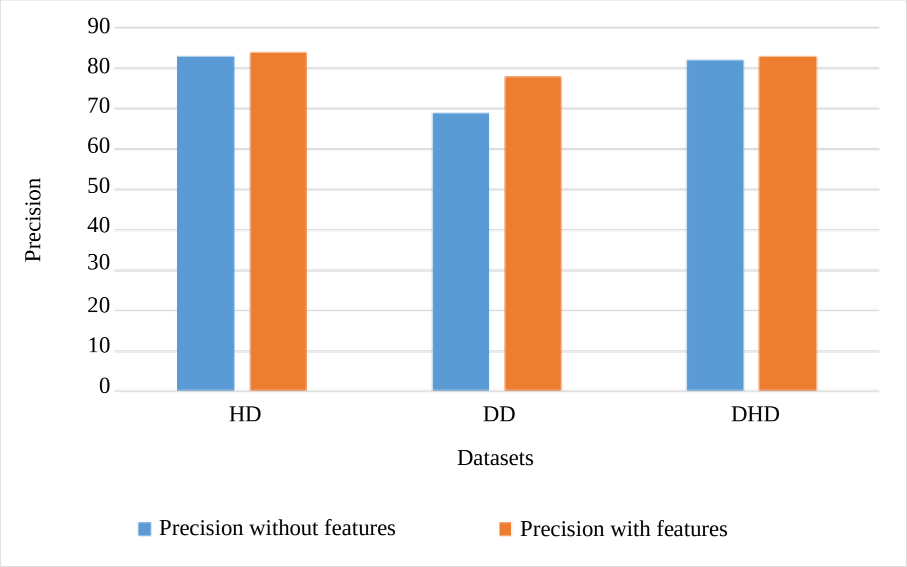

Fig. 6: Precision of RFS-RFE

From the graph, the precision values are increased for RFS-RFE by using important features for all the three datasets. The precision values are nearly achieved same such as 83% for without features and 84% for with features on HD datasets, whereas 82% for non-selection features and 83% for selected features on DHD datasets. But, for DD datasets alone, the precision values show some variations like 69% for without feature method and 84% for feature selection method. Finally, the last parameter like specificity is validated for different datasets are tabulated in Table 8 and Fig. 7.

Table 8: Specificity of RFS-RFE

	Specificity	
Datasets	Without features	With features
HD	91.54	92.66
DD	84.11	87.80
DHD	91.30	93.70

From the above diagram, it is clearly shown that specificity values for feature selection method achieved higher performance than RFS-RFE without feature selection method. The specificity of RFS-RFE with selection method achieved 92.66% for HD, 87.8% for DD and 93.75% for DHD datasets, whereas the RFS-RFE without selection method achieved 91.54%, 84.11% and 91.3% for HD, DD and DHD datasets. When compared with all three datasets, the DD datasets achieved very less values for various parameters even for feature selection method. Due to the less number of features in DD datasets, the performance values will eventually reduce and once the features are reduced for finding the importance, again it will affect the performance. In DD dataset, the RFS-RFE without feature selection method achieved 75.03% accuracy, 42.85% sensitivity, 69% precision and 84.11% specificity, whereas the RFS-RFE feature selection method achieved 78.57% accuracy, 57.4% sensitivity, 78% precision and 87.8% specificity. Based on the risk factors alone, we are predicting the heart disease among the diabetic patients. The prediction includes classifying, the percentage of patients suffering from the heart disease.

Comparative Analysis of RFS-RFE

In this section, the effectiveness of RF classifier with feature selection is validated for all the three datasets by comparing with ANN with fuzzy, SVM and fuzzy with NN (Vivekanandan and Iyengar, 2017). The existing method fuzzy with NN is used to compare the effectiveness of classifier for HD dataset alone in terms of various parameters. For comparing the effectiveness of classifier, this method implements the ANN with fuzzy and SVM for all three datasets in terms of accuracy, precision, sensitivity and specificity which are tabulated in Table 9.

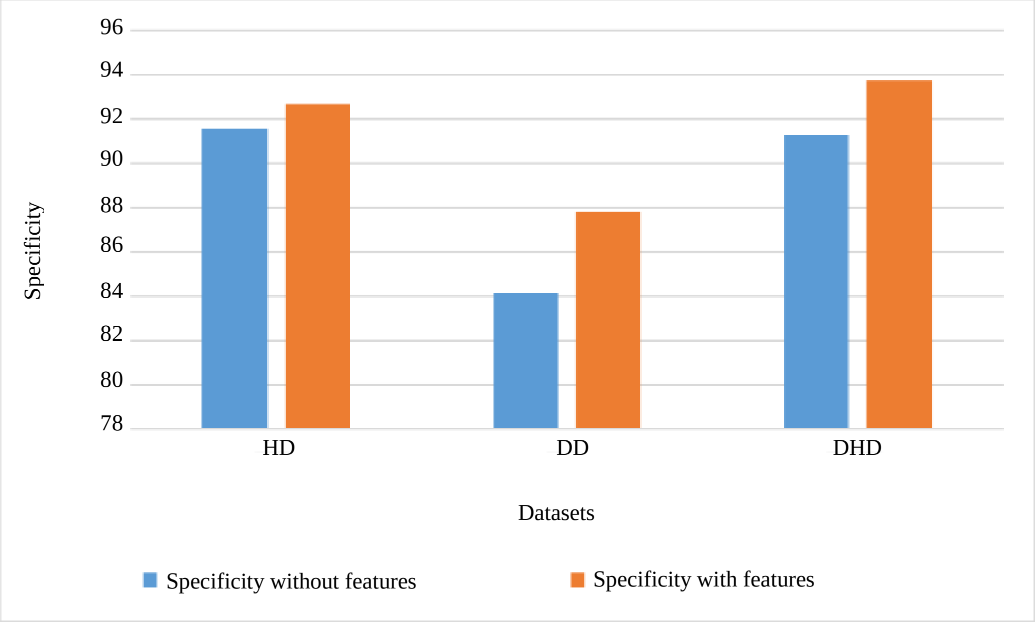

Fig. 7: Specificity for RFS-RFE

Table 9: Comparative study for feature selection methods

Datasets	Methods	Accuracy	Sensitivity	Specificity	Precision
HD	ANN with fuzzy	79.60	61.53	91.90	63
	SVM	79.70	63.15	92.15	81
	Fuzzy with NN (Vivekanandan and Iyengar, 2017)	83.00	84.00	89.00	74
	RFS-RFE	83.49	72.09	92.66	84
DD	ANN with fuzzy	68.39	83.53	88.53	65
	SVM	76.19	45.94	90.44	75
	RFS-RFE	78.57	57.40	87.80	78
DHD	ANN with fuzzy	79.81	66.60	91.06	80
	SVM	67.34	51.21	78.94	67
	RFS-RFE	82.56	66.66	93.75	83

The best values for this proposed dataset are highlighted in Table 6. The main objective of the proposed method is to find the risk factors of heart disease for diabetic patients by using important features. When compared with existing techniques Fuzzy with NN, the proposed RFS-RFE method achieved higher accuracy of 83.49 whereas the existing method achieved 83% of accuracy for HD datasets. Similarly, for the precision, sensitivity and specificity for DHD and HD dataset achieved better values. By using this method, the importance of features is calculated which is used to predict the risk factors of heart disease and can be avoided by the diabetic patients. The RFS-RFE method provides poor performance in DD dataset, when compared with other datasets for various parameters such as accuracy, precision, sensitivity and specificity. The reason for this degraded performance is due to the low number of features contained in DD dataset.

Conclusion

CV disease is one of the biggest cause for the highest death rate among the population of the world. There is a presence of a huge amount of raw data in the healthcare industry, hence prediction and decisions are made by data mining techniques using these collected raw data, which are transformed to information. The training and testing data is spliced into 70% and 30% after the preprocessing step using Data cleaning, Normalization etc., The training and testing data are utilized for feature extraction for selecting the most relevant features in the process to study and reduce the original high dimensional feature vector. In feature selection, still the presence of some unrelated data leads poor accuracy for prediction process. Therefore, the proposed RFE technique used to remove these irrelevant features from the selected data. In this work, the prediction of heart disease among diabetic patients using RFS feature selection. In addition, the output of RFE, which is the reduced subset of attributes, has been given as input to the Random Forest classifier to predict the heart disease among diabetic patients. The main advantage of the proposed system is that it considers all the important functions for the predicting the heart disease risk factors for diabetic patients and subsequently it is greater suitable for making recommendations based totally on deductive inference and prediction. While conducting the experiments on classifier, the RFS provides better performance because of minimal attributes sets, which is used to predict the heart disease. According to the contributions of critical attributes, the prediction accuracy is increased and also, the average prediction time is decreased by using the RFE algorithm, which is adopted for huge datasets. In future work, the performance of RFE will be further improved by using convolutional neural classifier for predicting the risk factor in higher efficiency.

Acknowledgement

We thank our Research and Innovation Cell, REVA University who provided insight and expertise that greatly assisted the research.

We thank them for assistance in process of implementation Dr. Vinayaka Murthy, REVA University for comments that greatly improved the manuscript.

We are also immensely grateful to peer reviewers for their comments on an earlier version of the manuscript, although any errors are our own and should not tarnish the reputations of these esteemed persons.

We are grateful for all the wonderful and resourceful authors for sharing their knowledge to all the researchers in the field.

We thank our families for their consistent support throughout the process of writing the paper.

Author’s Contributions

Bindushree Doddasiddavanahalli Channabasavaraju:

1. Authors make considerable contributions to conception and design and acquisition of data.
2. Analysis and interpretation of data and implementation.
3. Authors contribute in drafting the article.

Udayarani Vinayakamurthy:

1. Authors make considerable contributions to conception and design
2. reviewing it critically for significant intellectual content.
3. Gave final approval of the version to be submitted and any revised version.

Ethics

We would like to inform that the information stated here is genuine to the best of my knowledge.

The dataset which we have used is stated correctly.

We ensure that the references are cited as per the guidelines.

References

Amin, M.S., Y.K. Chiam and K.D. Varathan, 2019. Identification of significant features and data mining techniques in predicting heart disease. Telemat. Inform., 36: 82-93. DOI: 10.1016/j.tele.2018.11.007

Bhuvaneswari, G. and G. Manikandan, 2018. A novel machine learning framework for diagnosing the type 2 diabetics using temporal fuzzy ant miner decision tree classifier with temporal weighted genetic algorithm. Computing, 100: 759-772. DOI: 10.1007/s00607-018-0599-4

Bindushree, D.C., 2016. Prediction of cardiovascular risk analysis and performance evaluation using various data mining techniques: A review. Int. J. Eng. Res., 5013: 796-800.

Bindushree, D.C. and V.U. Rani, 2017. A review on using various DM techniques for evaluation of performance and analysis of heart disease prediction. Proceeding of the International Conference on Smart Technologies for Smart Nation, Aug. 17-19, IEEE Xplore Press, Bangalore, India, pp: 686-690. DOI: 10.1109/SmartTechCon.2017.8358459

Cui, S., D. Wang, Y. Wang, P.W. Yu and Y. Jin, 2018. An improved support vector machine-based diabetic readmission prediction. Comput. Meth. Programs Biomed., 166: 123-135. DOI: 10.1016/j.cmpb.2018.10.012

Chen, S.C., C.P. Lin, H.C. Hsu, J.H. Shu and Y. Liang *et al*., 2019. Serum bilirubin improves the risk predictions of cardiovascular and total death in diabetic patients. Clin. Chimica Acta, 488: 1-6. DOI: 10.1016/j.cca.2018.10.028

Guo, F. and W.T. Garvey, 2015. Development of a weighted Cardiometabolic Disease Staging (CMDS) system for the prediction of future diabetes. J. Clin. Endocrinol. Metab., 100: 3871-3877. DOI: 10.1210/jc.2015-2691

Hidalgo, J.I., J.M. Colmenar, G. Kronberger, S.M. Winkler and O. Garnica *et al*., 2017. Data based prediction of blood glucose concentrations using evolutionary methods. J. Med. Syst., 41: 142-142. DOI: 10.1007/s10916-017-0788-2

Kang, S., P. Kang, T. Ko, S. Cho and S.J. Rhee *et al*., 2015. An efficient and effective ensemble of support vector machines for anti-diabetic drug failure prediction. Expert Syst. Applic., 42: 4265-4273. DOI: 10.1016/j.eswa.2015.01.042

Kavitha, R. and E. Kannan, 2016. An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining. Proceeding of the International Conference on Emerging Trends in Engineering, Technology and Science, Feb. 24-26, IEEE Xplore Press, Pudukkottai, India. DOI: 10.1109/ICETETS.2016.7603000

Looker, H.C., M. Colombo, F. Agakov, T. Zeller and L. Groop *et al*., 2015. Protein biomarkers for the prediction of cardiovascular disease in type 2 diabetes. Diabetologia, 58: 1363-1371. DOI: 10.1007/s00125-015-3535-6

Long, N., P. Meesad and H. Unger, 2015. A highly accurate firefly based algorithm for heart disease prediction. Expert Syst. Applic., 42: 8221-8231. DOI: 10.1016/j.eswa.2015.06.024

Mathan, K., P.M. Kumar, P. Panchatcharam, G. Manogaran and R. Varadharajan, 2018. A novel Gini index decision tree data mining method with neural network classifiers for prediction of heart disease. Design Automat. Embedded Syst., 22: 225-242. DOI: 10.1007/s10617-018-9205-4

Mdhaffar, A., I.B. Rodriguez, K. Charfi, L. Abid and B. Freisleben, 2017. CEP4HFP: Complex event processing for heart failure prediction. IEEE Trans. Nanobiosci., 16: 708-717. DOI: 10.1109/TNB.2017.2769671

Maji, S. and S. Arora, 2019. Decision tree Algorithms for Prediction of Heart Disease. In: Information and Communication Technology for Competitive Strategies, Fong S., S. Akashe, P. Mahalle (Eds.), Springer, Singapore, ISBN-10: 978-981-13-0586-3 pp: 447-454.

Novara, C., N.M. Pour, T. Vincent and G. Grassi, 2016. A nonlinear blind identification approach to modeling of diabetic patients. IEEE Trans. Control Syst. Technol., 24: 1092-1100. DOI: 10.1109/TCST.2015.2462734

Ritika, C. and S. Mayank, 2016. Prediction of heart disease using data mining techniques. CSI Trans. ICT, 4: 193-198. DOI: 10.1007/s40012-016-0121-0

Sharma, P. and K. Saxena, 2017. Application of fuzzy logic and genetic algorithm in heart disease risk level prediction. Int. J. Syst. Assurance Eng. Manage., 8: 1109-1125. DOI: 10.1007/s13198-017-0578-8

Shouman, M., T. Turner and R. Stocker, 2013. Integrating clustering with different data mining techniques in the diagnosis of heart disease. J. Comput. Sci. Eng.

Prakash, S., K. Sangeetha and N. Ramkumar, 2018. An optimal criterion feature selection method for prediction and effective analysis of heart disease. Cluster Comput., 22: 1-7. DOI: 10.1007/s10586-017-1530-z

Vivekanandan, T. and N.C.S.N. Iyengar, 2017. Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease. Comput. Biol. Med., 90: 125-136. DOI: 10.1016/j.compbiomed.2017.09.011

Zhang, J., R.L. Lafta, X. Tao, Y. Li and F. Chen *et al*., 2017. Coupling a fast Fourier transformation with a machine learning ensemble model to support recommendations for heart disease patients in a tele health environment. IEEE Access, 5: 10674-10685. DOI: 10.1109/ACCESS.2017.2706318