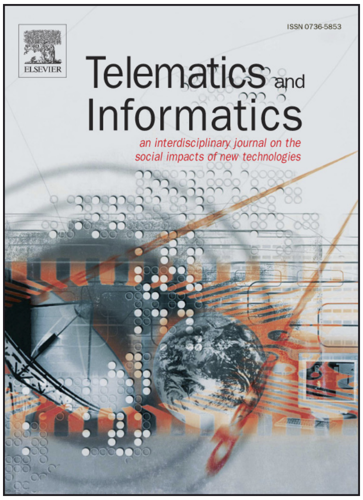


Identification of significant features and data mining techniques in predicting heart disease

Mohammad Shafenoor Amin, Yin Kia Chiam, Kasturi Dewi Varathan

PII: S0736-5853(18)30887-6
DOI: https://doi.org/10.1016/j.tele.2018.11.007
Reference: TELE 1191

To appear in: *Telematics and Informatics*

Received Date: 6 August 2018
Revised Date: 16 November 2018
Accepted Date: 20 November 2018

Please cite this article as: Shafenoor Amin, M., Kia Chiam, Y., Dewi Varathan, K., Identification of significant features and data mining techniques in predicting heart disease, *Telematics and Informatics* (2018), doi: https://doi.org/10.1016/j.tele.2018.11.007

Title: Identification of significant features and data mining techniques in predicting heart disease

Author names and affiliations:
Mohammad Shafenoor Amin[1,3], Yin Kia Chiam[1], Kasturi Dewi Varathan[2]
[1]Department of Software Engineering, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia
[2]Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia
[3]BRAC University, 66, Mohakhali, Dhaka 1212, Bangladesh.
Corresponding Authors:
Yin Kia Chiam
Department of Software Engineering, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia
e-mail: yinkia@um.edu.my

Declaration of interest: none

Title: Identification of significant features and data mining techniques in predicting heart disease.

Abstract

Cardiovascular disease is one of the biggest cause for morbidity and mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important subject in the section of clinical data analysis. The amount of data in the healthcare industry is huge. Data mining turns the large collection of raw healthcare data into information that can help to make informed decision and prediction. There are some existing studies that applied data mining techniques in heart disease prediction. Nonetheless, studies that have given attention towards the significant features that play a vital role in predicting cardiovascular disease are limited. It is crucial to select the correct combination of significant features that can improve the performance of the prediction models. This research aims to identify significant features and data mining techniques that can improve the accuracy of predicting cardiovascular disease. Prediction models were developed using different combination of features, and seven classification techniques: k-NN, Decision Tree, Naive Bayes, Logistic Regression (LR), Support Vector Machine (SVM), Neural Network and Vote (a hybrid technique with Naïve Bayes and Logistic Regression). Experiment results show that the heart disease prediction model developed using the identified significant features and the best-performing data mining technique (i.e. Vote) achieves an accuracy of 87.4% in heart disease prediction.

Keywords: Data mining; prediction model; classification algorithms; feature selection; heart disease prediction

Declaration of interest: none

1. Introduction

Cardiovascular disease (also known as heart disease) remains the number one cause of death throughout the world for the past decades. In 2015, the World Health Organization (WHO) has estimated that 17.7 million deaths have occurred worldwide due to cardiovascular diseases (WHO, 2017). CVDs are the number 1 cause of death globally: more people die annually from CVDs than from any other causes. If we can predict the cardiovascular disease and provide warning beforehand, a handful of deaths can be prevented.

The application of data mining brings a new dimension to cardiovascular disease prediction. Various data mining techniques are used for identifying and extracting useful information from the clinical dataset with minimal user inputs and efforts (Srinivas, Rani, & Govrdhan, 2010). Over the past decade, researchers explored various ways to implement data mining in healthcare in order to achieve an accurate prediction of cardiovascular diseases.

The efficiency of data mining largely varies on the techniques used and the features selected. The medical datasets in the healthcare industry are redundant and inconsistent. It is harder to use data mining techniques without prior and appropriate preparations. According to Kavitha and Kannan (2016), data redundancy and inconsistency in a raw dataset affect the predicted outcome of the algorithms. As a result, to apply the machine learning algorithms to its full potential, an effective preparation is needed to preprocess the datasets. Furthermore, unwanted features can reduce the performance of the data mining techniques as well (Paul et al., 2016). Thus, along with data preparation, a proper feature selection method is needed to achieve high accuracy in heart disease prediction using significant features and data mining techniques.

Although it has been quite clear that feature selection is as important as the selection of a suitable technique, researchers are still struggling in combining appropriate data mining technique with a proper set of features. According to Shouman et al. (2013), there is an expectation to diagnose the cardiovascular disease with high accuracy but it is not easy to achieve it. Additionally, a combination of significant features will definitely improve the accuracy of prediction. This shows that an extensive experiment to identify significant features is necessary to achieve that goal.

The performance of data mining techniques used in predicting cardiovascular disease is greatly reduced without a good combination of key features and also the improper use of the machine learning algorithms (Dey et al., 2016). Thus, it is crucial to identify the best combination of significant features that works incredibly well with the best performing algorithm. This research focuses on finding the data mining techniques with significant features that will perform well in predicting heart disease. However, it is not easy to identify the proper technique and select the significant features. Existing studies have shown that data mining techniques used in cardiovascular disease prediction are insufficient, and a proper examination is required to identify the significant features and data mining techniques that will improve the performance. According to Nahar et al. (2013), a proper evaluation and comparison to test the different combination of features together with the data mining techniques, are yet to be focused. Thus, a need for a thorough experimentation arises to provide proper identification of data mining techniques and significant features in order to ensure the prediction of heart disease is acceptable and accurate.

This research aims to identify the significant features and data mining techniques to predict heart disease. An experiment was conducted to identify the features and data mining techniques. The heart disease datasets were collected from the data source, UCI Machine Learning Repository. Cleveland dataset was selected because it is a commonly used database by machine learning researchers with records that are most complete. Seven classification techniques (k-NN, Decision Tree, Naïve Bayes, Logistic Regression, Vote, Support Vector Machine and Neural Network) were applied to create prediction models for this experiment using the prepared dataset. Based on the experiment results, nine significant features and top three data mining techniques were identified. The experiment results were evaluated using another dataset, UCI Statlog Heart disease dataset to confirm the findings. Additionally, this research also compares the highest accuracy achieved by the best technique identified from this research against the highest accuracy achieved in the existing studies.

The rest of the paper is organized as follows. Section 2 describes the heart disease dataset used in this research, to identify the significant features and data mining techniques. Section 3 explains the methods used to conduct the experiment that includes Data Preprocessing, Feature Selection, Classification Modelling using Data Mining Technique and Performance

Measure. The process of feature engineering is described to illustrate the selection of significant features in heart disease prediction. Section 4 presents the results of the experiment, which is the performance evaluation of the model created using seven data mining techniques. Section 5 discusses the analysis conducted to identify significant features and data mining techniques in order to create the best performing model. Section 6 describes the evaluation experiment conducted to validate the findings using another dataset. Finally, the last section concludes the study and presents the future work.

2. Dataset

The heart disease data were collected from UCI machine learning repository (Dua and Karra, 2017). There are four databases (i.e. Cleveland, Hungary, Switzerland, and the VA Long Beach). The Cleveland database was selected for this research because it is a commonly used database by machine learning researchers with records that are most complete. The dataset contains 303 records. Although the Cleveland dataset has 76 attributes, the dataset provided in the repository only provides information for a subset of 14 attributes. The data source of the Cleveland dataset is Cleveland Clinic Foundation. Table 1 describes the description and type of attributes. There are 13 attributes that feature in heart disease prediction and one attribute serves as the output or the predicted attribute for the presence of heart disease in a patient.

The Cleveland dataset contains an attribute named ‘num’ to show the diagnosis of heart disease in patients on different scales, from 0 to 4. In this scenario, 0 represents the absence of heart disease and all the values from 1 to 4 represent patients with heart disease, where the scaling refers to the severity of the disease (4 being the highest). Figure 1 shows the distribution of ‘num’ attribute among the 303 records.

Table 1: Description of attributes from UCI Cleveland Dataset

Attribute	**Description**	**Type**
Age	Age of the patient in years	Numeric
Sex	Gender of the patient (1 for male and 0 for female)	Nominal
Cp	Chest pain type described with 4 values; Value 1: typical angina Value 2: atypical angina Value 3: non-anginal pain Value 4: asymptomatic	Nominal
Trestbps	Resting blood pressure (in mm/Hg on admission to the hospital)	Numeric
Chol	Serum cholesterol in mg/dl	Numeric
Fbs	Fasting blood sugar > 120 mg/dl; 1 if true and 0 if false	Nominal
Restecg	Resting electrocardiographic results in 3 values; Value 0: normal Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria	Nominal
Thalach	Maximum heart rate achieved	Numeric
Exang	Exercise induced angina (1 for yes and 0 for no)	Nominal
Oldpeak	ST depression induced by exercise relative to rest	Numeric
Slope	The slope of the peak exercise ST segment	Nominal

	Value 1: upsloping Value 2: flat Value 3: downsloping	
Ca	Number of major vessels (0-3) coloured by fluoroscopy	Numeric
Thal	The heart status described with 3 values Value 3: normal Value 6: fixed defect Value 7: reversable defect	Nominal
Num	It represents the diagnosis of heart disease with 5 values. 0 meaning absence, and 1 to 4 meaning presence of heart disease.	Nominal

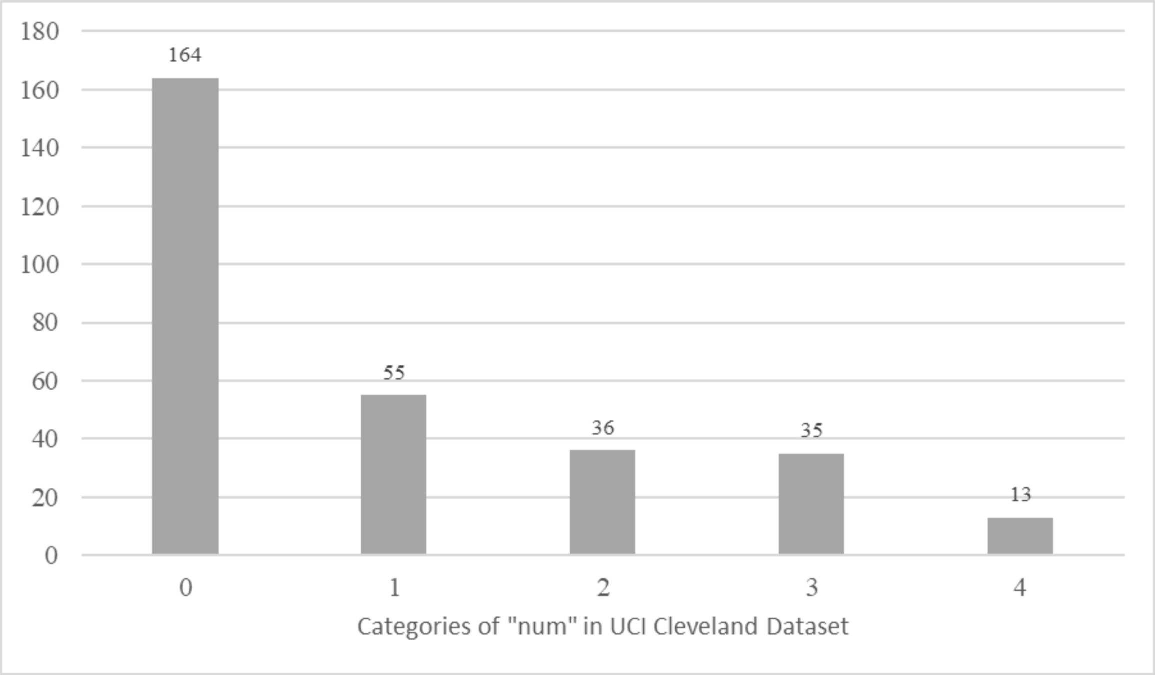


Figure 1: Distribution of "num" in UCI Cleveland dataset

3. Methods

In this research, RapidMiner Studio was used to conduct the experiment because it provides a robust and easy-to-use visual design environment for building of predictive analytic workflows. The visual representation of the workflow is one of the efficient feature for beginners. Moreover, it supports open source innovation, availability, and effective functionality. Figure 2 shows the workflow of the experiment. In the experiment, the UCI Cleveland heart disease dataset was imported into RapidMiner. The data mining process starts from the pre-processing phase, followed by feature engineering, to select different combination of attributes and classification modelling, to create models for prediction using data mining techniques. The feature selection and modelling were repeated for all the combination of the attributes. The loop iterates as a subset containing minimum 3 attributes that are chosen from the 13 attributes and the model is applied to it. The performance of each model created, based on the selected attributes and data mining technique during each iteration, is recorded and the output of the results is shown after the entire process is completed.

Section 3.1 to Section 3.4 describes in more detail the data preprocessing, feature selection, classification modelling, and performance measure. The results of performance measure are presented in Section 4.

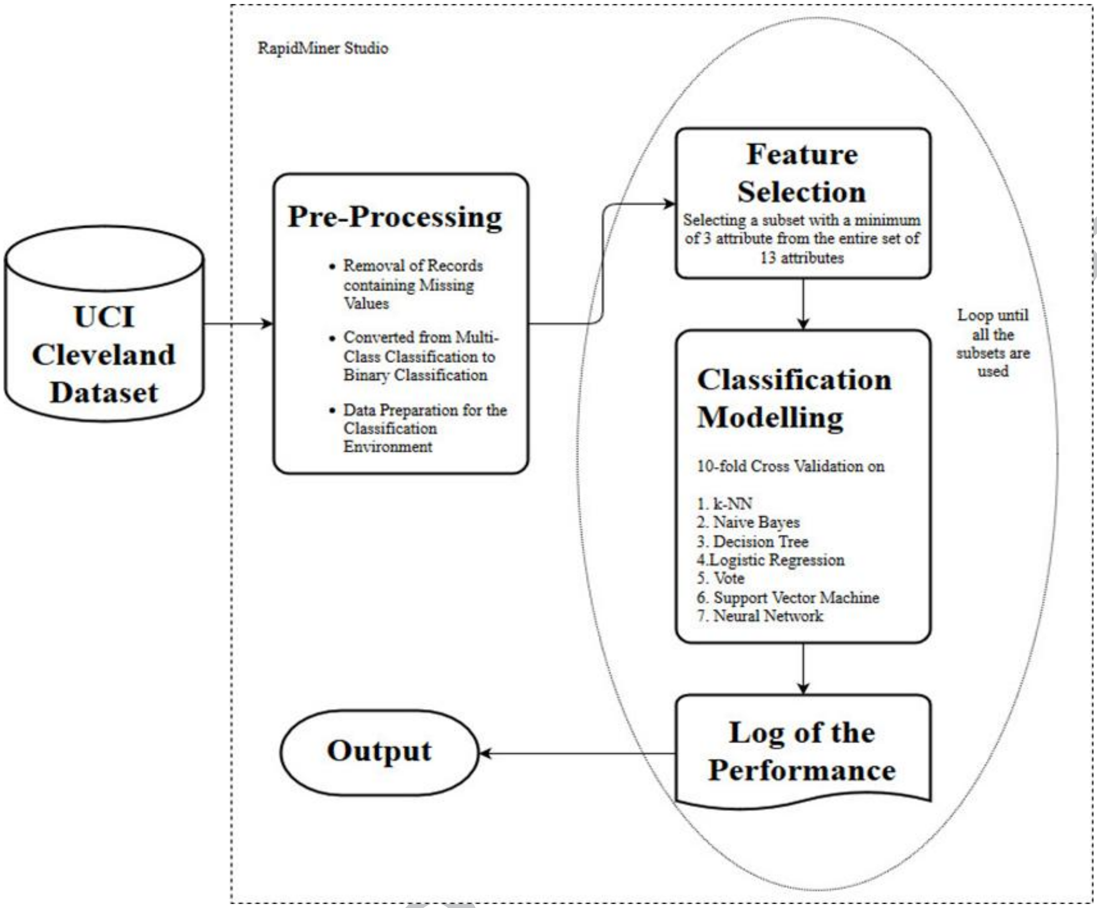

RapidMiner Studio
UCI Cleveland Dataset
Pre-Processing
Removal of Records containing Missing Values
Converted from Multi-Class Classification to Binary Classification
Data Preparation for the Classification Environment
Feature Selection
Selecting a subset with a minimum of 3 attribute from the entire set of 13 attributes
Classification Modelling
10-fold Cross Validation on
1. k-NN
2. Naive Bayes
3. Decision Tree
4.Logistic Regression
5. Vote
6. Support Vector Machine
7. Neural Network
Loop until all the subsets are used
Log of the Performance
Output

Figure 2: Workflow of experiment in this study

3.1 Data Preprocessing

The data were preprocessed after collection. There were 6 records that have missing values in Cleveland dataset. All the records with missing values were removed from the dataset, thus reducing the number of records from 303 to 297. 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 data preprocessing task was performed by converting all the diagnosis values from 2 to 4 into 1. The resulting dataset thus contains only 0 and 1 as the diagnosis value, where 0 being the absence and 1 being the presence of heart disease. After the reduction and transformation, the distribution of 297 records for ‘num’ attributes resulted in 160 records for ‘0’ and 137 records for ‘1’.

3.2 Feature Selection

Among the 13 features used in heart disease prediction, only ‘age’ and ‘sex’ attributes refer to the personal information of each patient. The remaining 11 features are all clinical attributes collected from various medical examinations. In this experiment, a combination of features were selected to be used with 7 classification techniques; k-NN, Decision Tree, Naïve Bayes, Logistic Regression, Vote, Support Vector Machine and Neural Network, to create the classification model. For this purpose, brute force method was applied to limit its lower bound (minimum 3 features). The procedure was to test each possible combination of features with all the techniques. In the experiment, firstly, all possible combinations of 3 features from the 13 attributes were chosen and each combination was tested by applying the 7 data mining techniques. Next, the experiment was repeated to select all possible combinations of 4 features from the 13 attributes.

The total number of combination achievable from a set of 13 attributes, excluding the empty set, is represented by $2^n - 1$. In this research, a single subset of the combination of features cannot have less than 3 attributes. Thus, all the subsets of combination achieved by having 2 attributes and 1 attribute are omitted. The equation used to calculate the total number of combinations is derived as follows.

Total number of combination

$$=\mathbf{2^n} - \left(\frac{\mathbf{n!}}{\mathbf{1!(n-1!)}}\right) - \left(\frac{\mathbf{n!}}{\mathbf{2!(n-2)!}}\right) - \mathbf{1}$$

$$= \mathbf{2^n} - \boldsymbol{n} - \frac{\mathbf{n\,(n-1)}}{\mathbf{2}} - \mathbf{1}$$

$$=\mathbf{2^n} - (\frac{\mathbf{2n+n^2-n}}{\mathbf{2}} + \mathbf{1})$$

$$=\mathbf{2^n} - (\frac{\mathbf{n^2+n}}{\mathbf{2}} + \mathbf{1})$$

where n represents the total number of features used to generate the subsets of combination, which is 13 in this experiment. Thus, a total of 8100 combinations of the features were selected and tested in this experiment.

3.3 Classification Modelling using Data Mining Technique

After selecting the features, the models were created with the 7 most popular classification techniques in data mining: k-NN, Decision Tree, Naive Bayes, Logistic Regression (LR), Support Vector Machine (SVM), Neural Network and Vote (i.e. a hybrid technique with Naïve Bayes and Logistic Regression). 10-folds cross validation technique was used to validate the performance of the models. In this technique, the entire dataset is divided into 10 subsets and then processed 10-times. 9 subsets are used as testing sets and the remaining 1 subset is used as a training set. Finally, the results are shown by averaging each 10 iterations. The subsets are divided using stratified sampling, meaning that each subset will have the same class ratio of the main dataset.

3.4 Performance Measure

The performance of the classification models was measured using three performance measures: accuracy, f-measure and precision. Accuracy is the percentage of correctly predicted instances among all instances. F-measure is the weighted mean of the precision and recall. Precision is the percentage of correct predictions for the positive class.

To identify the significant features, these three performance measures were used, whereas to identify data mining technique to create the best performing models, the accuracy and precision measures were used. For identification of significant features, the three performance measures give a better understanding of the overall behavior of the different combination of features. On the other hand, analysis of data mining techniques focusses on the best performing models that can produce high accuracy in heart disease prediction because accuracy and precision are the most intuitive evaluation metrics on performance. For each classifier, performances have been measured separately and all the results are recorded properly for further analysis.

4. Results

This section presents the results achieved in the experiments. Nine significant attributes: “sex”, “cp”, “fbs”, “restecg”, “exang”, “oldpeak”, “slope”, “ca” and “thal”, and the top three best performing techniques: Vote, Naïve Bayes and SVM, were identified based on the analysis of the experiment results. Section 4.1 shows the results of performance measure. Section 4.2 describes the analysis of features and data mining techniques.

4.1 Results of performance measure

The performance of 7 data mining techniques on 8100 combinations of features were experimented one by one. All the experiment results (including the accuracy, precision and f-measure of each model) were gathered for further analysis. Table 2, Table 3 and Table 4 describe the highest accuracy, the highest precision and the highest f-measure achieved by each data mining technique and the combination of features used in the model.

Table 2: The highest accuracy achieved by each data mining technique

Technique	Accuracy	Combination
Support Vector Machine (SVM)	86.87%	age, sex, cp, chol, fbs, exang, oldpeak, slope, ca
Vote	86.20%	sex, cp, fbs, thalach, exang, slope, ca, thal
Naïve Bayes	85.86%	sex, cp, thalach, exang, oldpeak, ca
Logistic Regression	85.86%	age, sex, cp, chol, restecg, oldpeak, slope, ca, thal
Neural Network	84.85%	sex, cp, trestbps, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal
k-NN	82.49%	sex, cp, fbs, restecg, oldpeak, ca, thal
Decision Tree	82.49%	sex, cp, fbs, restecg, oldpeak, ca, thal

Table 3: The highest precision achieved by each data mining technique

Technique	Precision	Combination
k-NN	95.00%	sex, restecg, exang
Decision Tree	95.00%	sex, restecg, exang
Vote	90.27%	cp, trestbps, fbs, thalach, exang, oldpeak, slope, ca, thal
Naïve Bayes	87.92%	sex, cp, fbs, slope, ca, thal
Support Vector Machine (SVM)	86.86%	sex, cp, chol, slope, ca
Neural Network	86.43%	sex, ca, thal
Logistic Regression	86.42%	sex, cp, trestbps, thalach, exang, oldpeak, slope, ca, thal

Table 4: The highest f-measure achieved by each data mining technique

Technique	F-measure	Combination
Support Vector Machine (SVM)	88.22%	age, sex, cp, chol, fbs, exang, oldpeak, slope, ca
Naïve Bayes	87.35%	sex, cp, thalach, exang, oldpeak, ca
Logistic Regression	87.27%	age, sex, cp, chol, restecg, oldpeak, slope, ca, thal
Neural Network	85.98%	sex, cp, trestbps, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal
Vote	84.41%	sex, cp, fbs, thalach, exang, slope, ca, thal
k-NN	84.05%	sex, cp, fbs, restecg, oldpeak, ca, thal
Decision Tree	84.05%	sex, cp, fbs, restecg, oldpeak, ca, thal

The three tables show the performance of 7 classification techniques in three different categories. Based on the analysis showed in the tables, we can see that the highest accuracy (86.87%) was achieved by SVM with 9 attributes. On the other hand, the highest precision (95.00%) was achieved by both Decision Tree and k-NN, using the same combination of 3 features (i.e. sex, restecg, exang) whereas the highest f-measure was given by SVM with 9 attributes. Table 2 shows the top three best performing techniques with an accuracy of more than 90%, which are SVM, Naïve Bayes and Vote. The results also indicate that both Decision Tree and k-NN have the lowest accuracy (82.49%) as compared to other techniques. However, these two techniques gave the highest performance in precision.

Table 5 shows the average accuracy achieved by each data mining technique on all of the 8100 combinations of features. Table 6 and Table 7 show the average precision and F-measure of each data mining technique. Based on Table 5, Vote, Naïve Bayes and SVM are the top three techniques for all of the 8100 combinations by getting an average of 78.20% and 78.15% in accuracy. On the other hand, the average values of precision shown in Table 6 indicate that Vote, Naïve Bayes and SVM are the top three techniques. According to Table 7, the top three techniques that have achieved the highest average F-measure are LR, SVM and Naïve Bayes.

Table 5: Average accuracy achieved by each data mining technique

Technique	**Average Accuracy Achieved**
Vote	78.20%
Naïve Bayes	78.20%
Support Vector Machine (SVM)	78.15%
Logistic Regression (LR)	78.03%
Neural Network	75.18%
k-NN	63.50%
Decision Tree	63.50%

Table 6: Average precision achieved by each data mining technique

Technique	Average Precision Achieved
Vote	79.41%
Naïve Bayes	78.76%
Support Vector Machine (SVM)	78.15%
Logistic Regression (LR)	76.27%
Neural Network	76.20%
k-NN	66.43%
Decision Tree	66.43%

Table 7: Average f-measure achieved by each data mining technique

Technique	Average F-measure Achieved
Logistic Regression (LR)	80.98%
Support Vector Machine (SVM)	80.25%
Naïve Bayes	80.17%
Vote	78.10%
Neural Network	77.33%
k-NN	65.87%
Decision Tree	65.87%

4.2 Analysis of Features and Data Mining Technique

This section describes the selection of significant features and data mining techniques based on the results obtained from the experiments. By analyzing these results, the significant features and data mining technique that have significant impacts in creating the best performing models are identified to predict heart disease. Section 4.2.1 and Section 4.2.2 show the analysis of the significant features, and the best performing data mining techniques selected in this research.

4.2.1 Feature Selection

The results achieved from the experiments were further analyzed to identify the significant attributes to predict the presence of heart disease. In order to identify the significant attributes, an analysis was conducted to find out how many times an attribute was selected in the model that has performed with the highest accuracy, precision and F-measure. Among all of the 8100 combinations, the combination of features that has resulted in the highest performance of a specific technique was identified. Table 8 shows the analysis of attributes that have achieved the best performances on all of the data mining techniques. Thus, 7 techniques that have different combinations resulting in the highest performances. In this table, an attribute that has occurred in those highest performing combination has been counted and compared with the other attributes. The first row of Table 8 depicts how many times each of those attributes is found among the combinations that have resulted in the highest accuracy among the 7 techniques. Similarly, the second and third row depict the occurrences of the attributes which gave the highest performing F-measure and precision. Lastly, a summation of all the occurrences of each attribute are calculated.

Among all of the 13 attributes, 'sex' is the attribute with the highest number of total occurrence, appearing 21 times in all of the combinations. This indicates that this attribute is the most significant attribute that has an impact on predictions with high accuracy, F-measure and precision. In this research, the attributes that have appeared for at least 10 times, resulting with the highest performance, are identified as significant features in heart disease prediction. Based on the analysis in Table 8, nine attributes are identified as significant features in heart prediction: "sex", "cp", "fbs", "restecg", "exang", "oldpeak", "slope", "ca" and "thal". Furthermore, based on the experiment results, these nine attributes have been used in the highest accuracy models created using four or more data mining techniques.

Table 8: Comparison between Attributes resulting in the highest performance

Features Occurrence	Age	Sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
Occurrence in the Highest Accuracy	2	7	7	1	2	5	4	3	4	6	4	7	5
Occurrence in the Highest F-measure	2	7	7	1	2	5	4	3	4	6	4	7	5
Occurrence in the Highest Precision	0	6	4	2	1	2	2	2	4	2	4	5	4
Total Number of Occurrence	4	20	18	4	5	12	10	8	12	14	12	19	14

4.2.2 Selection of Data Mining Techniques

For the completion of our proposed model, we required the data mining technique to go hand in hand with the selected significant attributes. In this research, the top three data mining techniques are identified according to the highest average accuracy and precision obtained from the experiments. According to Table 5, the top three best performing data mining techniques in terms of the highest average accuracy and precision are selected. These three techniques are Vote, Naïve Bayes and SVM. In order to finalize the choice of the top three techniques, the results were cross-checked with the results obtained in Table 2 and Table 3. The comparison of the results show that these three techniques have appeared either as one of the top three or top four techniques in the highest accuracy and precision. Thus, Vote, Naïve Bayes and SVM are selected to develop the heart disease prediction models.

5. Evaluation

In this study, an evaluation was conducted to validate the findings of the significant features and data mining techniques identified in Section 4. Three prediction models were created using the nine significant attributes and the top three data mining techniques. Experiment was performed using another dataset, UCI Statlog Heart Disease dataset to validate the performance of the prediction models. Same as the Cleveland dataset, this dataset was collected from UCI machine learning repository (Dua and Karra, 2017). The structure of the Statlog heart disease dataset is similar to the Cleveland heart disease dataset. Table 9 shows the comparison between Statlog and Cleveland datasets. Both datasets have 13 attributes that feature the heart disease and 1 predicted attribute to show the presence of heart disease. All the names of the attributes are the same. The only difference in the attributes between the two datasets is the value that they used to represent the class attribute, "num". The output for the Statlog dataset contains two values: 1 and 2. "1" is the absence of heart disease and "2" is the presence of heart disease in patients. On the other hand, the Cleveland dataset has five different levels of "num" scaling from 0 to 4.

In the previous experiment, Cleveland dataset was converted from the multiclass values for 'num' into binary values (0, 1). This overcomes the difference of 'num' values between the two datasets. After preprocessing the Cleveland data, the distribution of 297 records for 'num' attributes resulted in 160 records for '0' and 137 records for '1'. On the other hand, as seen in Table 9, the Statlog dataset contains a total of 270 records. The dataset does not contain any missing values. The distribution of "1" and "2" as the value of "num" is 150 and 120 respectively. Since there is no missing values, the dataset did not require much preprocessing before the data is used in this evaluation. Furthermore, the ratio of the number of records for absence and presence of heart disease is almost the same (i.e. Cleveland’s ratio =160: 137; Statlog’s ratio = 150: 120).

Statlog dataset has a very clean representation of the records, thus increasing its popularity among the researchers as well. Many recent studies (Subbulakshmi et al., 2015, Nahato et al., 2015, Bashir et al., 2015, Srinivas et al., 2015) have used Statlog dataset in their experiments. Due to all of the similarity and quality of the data, the Statlog dataset was identified as the best dataset to be used in the evaluation to validate the proposed significant attributes and data mining techniques.

Table 9: Comparison between Cleveland and Statlog datasets

Comparison Category	**Cleveland Dataset**					**Statlog Dataset**	
No. of Attributes	13					13	
Attributes	age, sex, cp, trestbps, chol, fbs, restecg, exang, oldpeak, slope, ca, thal					age, sex, cp, trestbps, chol, fbs, restecg, exang, oldpeak, slope, ca, thal	
Class Attribute	num					num	
Different values for “num”	0,1,2,3,4					1,2	
Distribution of “num”	0	1	2	3	4	1	2
	164	55	36	35	13	150	120
Records with Missing Values	6					0	
Total number of instances	303					270	

5.1 Experimental setup for evaluation

Figure 3 shows the entire process used to conduct the experiment for evaluation. First, the Statlog data were preprocessed before using the dataset for evaluation. To make the Statlog dataset similar to the Cleveland dataset, class value of “num” was converted from “1” to “0” and from “2” to “1”. The resulting dataset thus contained 0 and 1 as the predicted output values where 0 is the absence and 1 is the presence of heart disease. After the transformation, the distribution of 270 records becomes 150 instances for ‘0’ and 120 instances for ‘1’. The data was then ready to be used for the classification environment.

The subset of nine significant features identified (i.e. sex, cp, fbs, restecg, exang, oldpeak, slope, ca and thal) were selected from the preprocessed dataset. Next, the classification models were developed using the top 3 data mining techniques (i.e. Vote, Naive Bayes and Support Vector Machine). In this experiment, 10-folds cross validation technique was used to measure the performance of the models. Finally, the accuracy results are shown by averaging the results obtained from the 10 iterations.

The evaluation of the model was performed with the help of confusion matrix. There are four outcomes based on confusion matrix: True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). The following formula was used to measure the accuracy of the classification models: Accuracy = (TP+TN)/n, where n = Total Number of Instances (Powers & Martin, 2011). Accuracy was selected as the performance measure because it is one of the most popular and intuitive criterion used in many existing studies to compare the performance of the classification models.

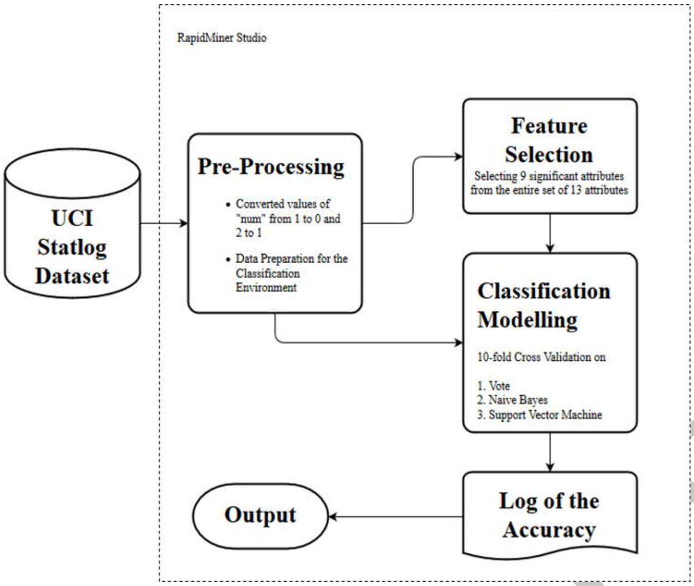

RapidMiner Studio
UCI Statlog Dataset
Pre-Processing
• Converted values of "num" from 1 to 0 and 2 to 1
• Data Preparation for the Classification Environment
Feature Selection
Selecting 9 significant attributes from the entire set of 13 attributes
Classification Modelling
10-fold Cross Validation on
1. Vote
2. Naive Bayes
3. Support Vector Machine
Log of the Accuracy
Output

Figure 3: Workflow of evaluation experiment in this study

5.2 Evaluation Results

The accuracy results achieved by the prediction models developed using the 9 significant features and top three classification modelling techniques are as follows:

- Accuracy of Vote with 13-attribute dataset= (103+130)/270=236/270=0.8630
- Accuracy of NB with 13-attribute dataset = (97+130)/270=236/270= 0.8407
- Accuracy of SVM with 13-attribute dataset = (89+133)/270=236/270=0.8222
- Accuracy of Vote with 9-attribute dataset = (100+136)/270=236/270=0.8741
- Accuracy of NB with 9-attribute dataset = (97+132)/270=236/270=0.8481
- Accuracy of SVM with 9-attribute dataset = (94+136)/270=236/270=0.8519

Table 10 summarizes the accuracy of the models obtained from the experiment. This table compares the accuracy of classification models for 13 attributes and 9 significant attributes. Based on Table 10, the accuracy of prediction models developed using the 9 significant features is better than the models developed using all 13 attributes. The highest accuracy (86.30%) for the

13-feature classification model was achieved by Vote. Additionally, the highest accuracy (87.41%) for the 9-feature classification model was also achieved by Vote.

Table 10: Accuracy obtained from the evaluation experiment

	Vote	Naïve Bayes	Support Vector Machine
Accuracy obtained with 13 features	86.30%	84.07%	82.22%
Accuracy obtained with identified 9 significant features	**87.41%**	84.81%	85.19%

6. Discussion

The results presented in Table 10 indicate that the identified significant features have improved the accuracy of all the top three data mining techniques. This confirms the findings presented in Section 4 on the significant attributes in heart disease prediction. Among the 9 features, 8 out of them are clinical attributes collected from various medical examinations. Only one feature, sex, is the attribute related to the demographic of a patient. This show that attributes on clinical investigation and examination have higher impact than information on demographic in predicting heart disease using data mining techniques.

According to Table 10, the prediction model developed using the hybrid data mining technique, Vote, and 9 significant features has achieved the highest accuracy of 87.41%. Since Vote outperforms the other two techniques in the second experiment and shows consistent accuracy in both experiments, it (Vote) was identified as the best performing technique among the top three techniques. The findings have encouraged further research to explore hybrid data mining techniques using different combination of data mining techniques to improve the performance of the prediction models.

Based on the evaluation results, the best prediction model was proposed in this research using the 9 significant attributes (sex, cp, fbs, restecg, exang, oldpeak, slope, ca and thal) and the Vote hybrid technique. Figure 5 shows the overview of the proposed heart disease prediction model. In addition to this, a heart disease prediction system was developed based on the proposed model to help to predict the presence of heart disease in a patient automatically. Figure 5 illustrates a screenshot of the heart disease prediction system. This system (HDPS v2.0) can be downloaded via this website: https://sites.google.com/um.edu.my/ykchiam/research/healthcare-data-analytics.

In overall, this study demonstrates that the identified significant features and data mining techniques have increased the performance of the prediction models. The proposed prediction model pave the way for further research on cardiovascular disease prediction that can support the decision making of clinicians in diagnosing patients with heart disease.

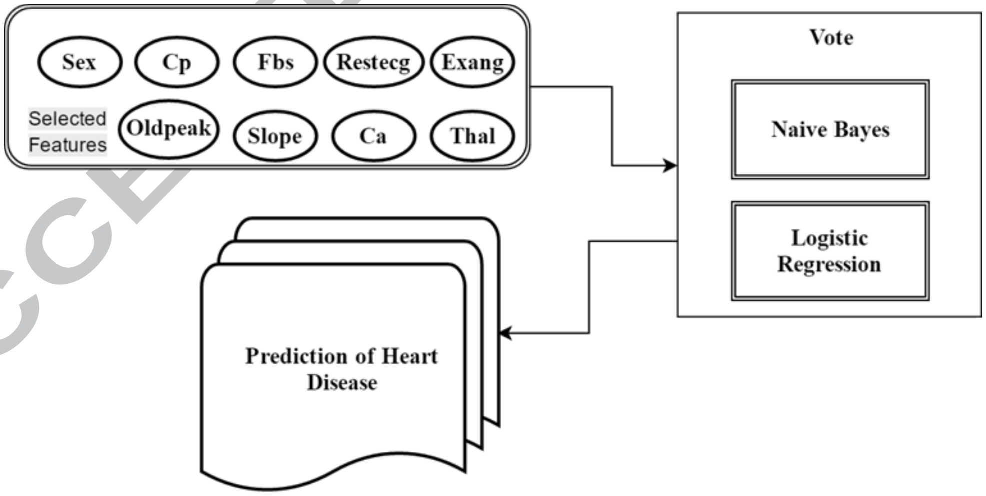
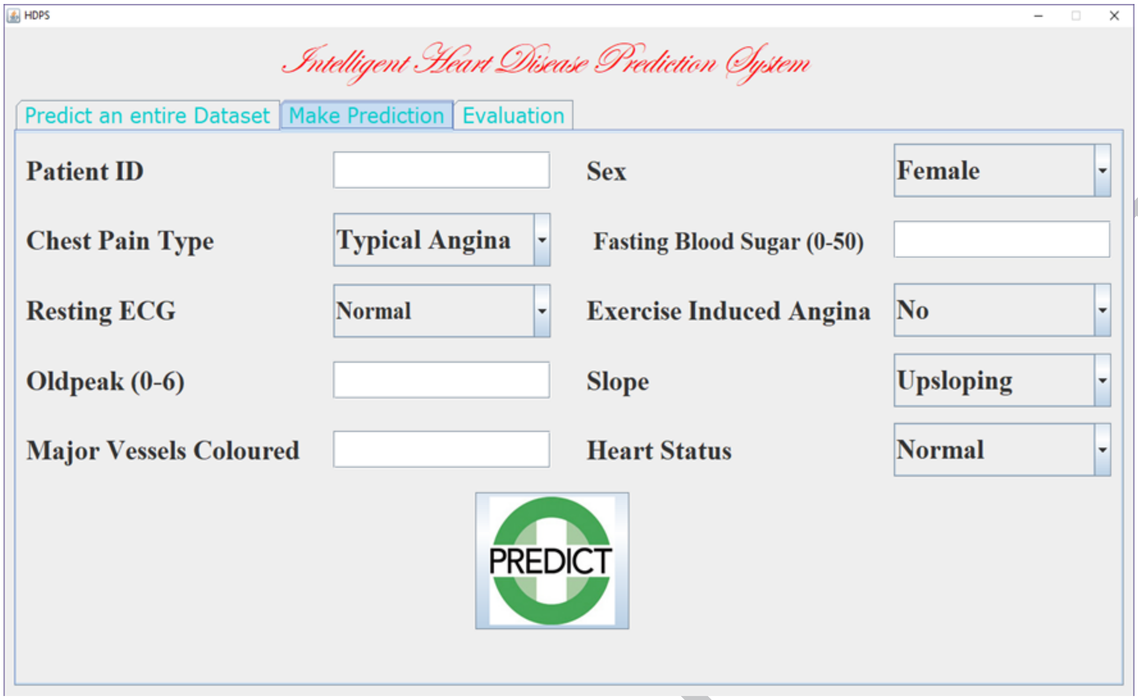


Figure 4: Proposed prediction model developed using Vote and nine significant features



Intelligent Heart Disease Prediction System

Patient ID		Sex	Female
Chest Pain Type	Typical Angina	Fasting Blood Sugar (0-50)	
Resting ECG	Normal	Exercise Induced Angina	No
Oldpeak (0-6)		Slope	Upsloping
Major Vessels Coloured		Heart Status	Normal

Figure 5: Screenshot of the heart disease prediction system developed based on the proposed prediction model

6.1 Benchmarking of the Proposed Model

Benchmarking is useful to compare the performance of a model against the performance achieved by other models. This method was used to assess whether the proposed model has achieved acceptable accuracy as compared to the accuracy achieved by other studies. The accuracy of the proposed model using the nine significant features (i.e. sex, cp, fbs, restecg, exang, oldpeak, slope, ca, thal) and Vote technique was benchmarked against the other six studies that have been conducted in recent years using UCI machine learning repository.

Table 11 shows the benchmarking of the accuracy of the proposed model against the accuracy of the models reported in the six existing studies. Based on Table 11, we can see that the proposed model has performed better as compared to the existing studies. Based on the comparison, it is apparent that this research has generated higher accuracy using the hybrid technique, Vote. Additionally, the classification model proposed in this research has proved to have acceptable accuracy and perform better than the other studies.

Table 11: Benchmarking of the Proposed Model

Source	**Technique Used**	**Accuracy Achieved**
The proposed model	Vote with Naïve Bayes and Logistic Regression	**87.41%**
Paul, Shill, Rabin, & Akhand (2016)	Neural Network with Fuzzy	80%
Verma, Srivastava & Negi (2016)	Decision Tree	80.68%
Ismaeel, Miri, Sadeghian & Chourishi (2015)	Extreme Learning Machine	86.50%
El-Bialy, Salamay, Karam & Khalifa (2015)	Decision Tree	78.54%
Subanya & Rajalaxmi (2014)	SVM	86.76%
Nahar, Imam, Tickle & Chen (2013)	Naïve Bayes	69.11%
Chaurasia & Pal (2013)	CART	83.49%
Khemphila & Boonjing (2011)	Neural Network with Genetic Algorithm	80.99%
Shouman, Turner & Stocker (2011)	Decision Tree with Gain Ratio	84.10%

Table 12 shows the features selected by twelve existing studies that have used UCI heart disease dataset for developing prediction models. The nine features selected in the proposed model have been used in at least five studies. This proves that the features selected in this study is significant for predicting the presence of heart disease in a patient.

Table 12: Features selected by studies using UCI heart disease dataset

Source	**sex**	**cp**	**fbs**	**restecg**	**exang**	**oldpeak**	**slope**	**ca**	**thal**
Liu, Wang, Su, Zhang, Zhu, Wang & Wang (2017)		√		√			√	√	√
Wiharto, Kusnanto & Herianto (2017)	√	√			√	√	√		
Dey, Singh & Singh (2016)	√	√	√	√	√		√		

Paul, Shill, Rabin & Akhand (2016)	√	√		√	√	√	√	√	√
Nahato, Harichandran & Arputharaj (2015)		√		√		√		√	√
Tomar & Agarwal (2014)		√	√	√		√	√	√	
Chaurasia & Pal (2013)	√	√	√	√	√		√		
Sen, Patel & Shukla (2013)	√	√	√	√	√				
Nahar, Imam, Tickle & Chen (2013)		√	√	√	√				
Bhatla & Jyoti (2012)		√			√	√		√	√
Anooj (2012)						√			√
Khemphila & Boonjing (2011)		√			√	√	√	√	√
Total	**5**	**11**	**5**	**8**	**8**	**7**	**7**	**6**	**6**

7. Conclusions

The clinical industry has a huge patient data that are not processed. Finding a way to process this raw data into a gem of information can save a lot of life. Data mining techniques can be used to analyze the raw data, to provide new insights towards the goal of disease prevention with accurate predictions. Heart disease is one of the main causes of deaths in the world. It is crucial to detect it in patients as soon as possible to prevent fatality.

In this study, significant features and the best performing classification modelling techniques that improve the accuracy of heart disease prediction were selected. An experiment was first conducted using the UCI Cleveland dataset to identify the significant features and the top three data mining techniques. The findings were evaluated through another experiment using UCI Statlog dataset. The nine significant features selected in this research are sex, cp, fbs, restecg, exang, oldpeak, slope, ca and thal. The top three data mining techniques that produce high accuracy in prediction are identified in this research as Vote, Naïve Bayes and Support Vector Machine. The evaluation results reconfirm that the nine selected features are significant.

Additionally, among the top three techniques, Vote has outperformed the other two techniques. The best prediction model was created using the nine significant attributes and the Vote technique. Finally, the accuracy of the proposed model was benchmarked against the accuracy of the models proposed in the existing studies. The outcome of the benchmarking indicates that the proposed classification model has produced a higher accuracy in prediction and performed better than the other studies.

There are many ways to enhance this research and address the limitations of this study. This research can be extended by conducting the same experiment on a large-scale real-world dataset. The Vote technique used in the proposed model is a hybrid technique that combines Naïve Bayes and Logistic Regression. Further research can be conducted to test different combination of data mining techniques in heart disease prediction. Additionally, new feature selection methods can be applied to get a broader perspective on the significant features to improve the accuracy in prediction.

Acknowledgments

This work was supported by the University of Malaya Research Grant (UMRG), Project Code: RP028C-14HTM and Fundamental Research Grant Scheme (FRGS), Project Code: FP057-2017A.

References

Anbarasi, M., Anupriya, E., & Iyengar, N. C. S. N. 2010. Enhanced prediction of heart disease with feature subset selection using genetic algorithm. International Journal of Engineering Science and Technology,2(10), 5370-5376.

Anooj, P. K. 2012. Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. Journal of King Saud University-Computer and Information Sciences, 24(1), 27-40.

Bhatla, N., & Jyoti, K. 2012. An analysis of heart disease prediction using different data mining techniques. International Journal of Engineering, 1(8), 1-4.

Chaurasia, V., Pal, S., 2013. Early prediction of heart diseases using data mining techniques. Carib. J. SciTech. 1, 208-217.

Dey, A., Singh, J., Singh, N., 2016. Analysis of Supervised Machine Learning Algorithms for Heart Disease Prediction with Reduced Number of Attributes using Principal Component Analysis. Analysis. 140(2), 27-31.

Dua, D., Karra Taniskidou, E., 2017. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. http://archive.ics.uci.edu/ml

El-Bialy, R., Salamay, M. A., Karam, O. H., Khalifa, M. E., 2015. Feature analysis of coronary artery heart disease data sets. Procedia Computer Science. 65, 459-468.

Ismaeel, S., Miri, A., Sadeghian, A., Chourishi, D., 2015. An Extreme Learning Machine (ELM) Predictor for Electric Arc Furnaces' vi Characteristics. IEEE 2nd International Conference on Cyber Security and Cloud Computing (CSCloud), New York, pp. 329-334.

Kavitha, R., & Kannan, E., 2016. An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining. International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS), Pudukkottai, pp. 1-5.

Khemphila, A., & Boonjing, V. 2011. Heart disease classification using neural network and feature selection. In 21st International Conference on Systems Engineering (ICSEng), Las Vegas, pp. 406-409. IEEE.

Liu, X., Wang, X., Su, Q., Zhang, M., Zhu, Y., Wang, Q., & Wang, Q. 2017. A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method. Computational and mathematical methods in medicine, 2017.

Nahar, J., Imam, T., Tickle, K. S., & Chen, Y. P. P., 2013. Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. Expert Systems with Applications. 40(1), 96-104.

Nahato, K. B., Harichandran, K. N., & Arputharaj, K. 2015. Knowledge mining from clinical datasets using rough sets and backpropagation neural network. Computational and Mathematical Methods in Medicine, 2015, 1-13.

Paul, A. K., Shill, P. C., Rabin, M. R. I., & Akhand, M. A. H. 2016. Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease. In 5th International Conference on Informatics, Electronics and Vision (ICIEV), pp. 145-150. IEEE.

Sen, A. K., Patel, S. B., & Shukla, D. D. 2013. A data mining technique for prediction of coronary heart disease using neuro-fuzzy integrated approach two level. International Journal Of Engineering And Computer Science (IJECS), 2(8), 2663-2671.

Shouman, M., Turner, T., Stocker, R. 2011. Using decision tree for diagnosing heart disease patients. Proceedings of the Ninth Australasian Data Mining Conference (AusDM’11), Darlinghurst, Australia, Volume 121, pp. 23-30.

Shouman, M., Turner, T., Stocker, R., 2013. Integrating clustering with different data mining techniques in the diagnosis of heart disease. J. Comput. Sci. Eng. 20(1).

Srinivas, K., Rani, B. K., Govrdhan, A., 2010. Applications of data mining techniques in healthcare and prediction of heart attacks. International Journal on Computer Science and Engineering (IJCSE). 2(02), 250-255.

Subanya, B., & Rajalaxmi, R. R., 2014. Feature selection using Artificial Bee Colony for cardiovascular disease classification. International Conference on Electronics and Communication Systems (ICECS), Coimbatore, pp. 1-6.

Tomar, D., & Agarwal, S. 2014. Feature selection based least square twin support vector machine for diagnosis of heart disease. International Journal of Bio-Science and Bio-Technology, 6(2), 69-82.

Verma, L., Srivastava, S., & Negi, P. C., 2016. A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data.Journal of Medical Systems. 40(7), 1-7.

Wiharto, W., Kusnanto, H., & Herianto, H. 2017. Hybrid system of tiered multivariate analysis and artificial neural network for coronary heart disease diagnosis. International Journal of Electrical and Computer Engineering (IJECE), 7(2), 1023-1031.

World Health Organization (WHO), 2017. Cardiovascular diseases (CVDs) – Key Facts. http://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

Highlights (3 to 5 bullet points, maximum 85 characters):

- Evaluation of seven classification algorithms in predicting heart disease.
- Experiments to select significant features and data mining techniques for heart disease prediction.
- The performance of classification models for heart disease prediction is investigated.
- A prediction model is developed using significant features and hybrid data mining technique.