International Journal on Advanced Science Engineering Information Technology

Vol.8 (2018) No. 6
ISSN: 2088-5334

Performances Analysis of Heart Disease Dataset using Different Data Mining Classifications

Wan Hajarul Asikin Wan Zunaidi[1], RD Rohmat Saedudin[2], Zuraini Ali Shah[1*], Shahreen Kasim[3], Choon Sen Seah[3] and Maman Abdurohman[4],

[1]*Faculty of Computing, Universiti Teknologi Malaysia, Johor, Malaysia*
**E-mail: aszuraini@utm.my*

[2]*School of Industrial Engineering, Telkom University, 40257 Bandung, West Java, Indonesia*

[3]*Faculty of Computer Sciences and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Malaysia*

[4]*School of Computing, Telkom University, 40257 Bandung, West Java, Indonesia*

***Abstract*— nowadays, heart disease is one of the major diseases that cause death. It is a matter for us to concern in today's highly chaotic life style that leads to various diseases. Early prediction of identification to heart-related diseases has been investigated by many researchers. The death rate can be further brought down if we can predict or identify the heart disease earlier. There are many studies that explore the different classification algorithms for classification and prediction of heart disease. This research studied the prediction of heart disease by using five different techniques in WEKA tools by using the input attributes of the dataset. This research used 13 attributes, such as sex, blood pressure, cholesterol and other medical terms to detect the likelihood of a patient getting heart disease. The classification techniques, namely J48, Decision Stump, Random Forest, Sequential Minimal Optimization (SMO), and Multilayer Perceptron used to analyze the heart disease. Performance measurement for this study are the accuracy of correct classification, mean absolute error and kappa statistics of the classifier. The result shows that Multilayer Perceptron Neural Networks is the most suited for early prediction of heart diseases.**

***Keywords*— WEKA; data mining; attribute selection; classification; heart disease.**

I. INTRODUCTION

Healthcare industry has focused on analyzing data to diagnose patient's diseases. Heart disease is one of the major diseases that focuses by the healthcare industry. By analyzing the patient's data, the quality of diagnosis can be improved. Proper diagnosis can cure the disease by providing the right treatment, while poor diagnosis can lead to wrong treatment which is unacceptable consequences. According to a survey of WHO (World Health Organization), heart attacks and strokes are the major cause to the 17 million total global deaths. Generally, the issues that are caused heart disease are mental stress, imbalance workload, and nutrition. Overall, it is very common to happen among adults.

Several different research groups have been studied that related to heart diseases using classifications techniques [1]–[4]. The studies that have been done were included the study of heart disease classification using many classifications algorithm. In this research, we have used five classification techniques. Classification is the process of finding a set of models that differentiate the data by telling the classes and meaning behind the data. The models are used to predict the class whose label is unknown. Famous classification algorithm such as decision trees, neural networks, Naïve Bayes, and SMO are being studied and further compare the usability of this research. In this research, we focused on heart disease classification using five classifiers.

Heart disease is studied, diagnosed and prepared as dataset by biologists. Researchers have then used the dataset and applied to different classification techniques for diagnosis and

achieved different probabilities for different methods. One of the intelligent healthcare systems which utilized the big data and data mining tools to perform the heart disease prediction was proposed by Palaniappan [5]. In his research, three techniques, Decision Trees, Naïve Bayes, and Neural Network are studied and applied using CRISP-DM methodology to conduct a better prediction system for heart disease. Throughout his study, the outcome of the Neural Network is better compared to other techniques. Overall, he introduced a supervised network which can help in heart diseases diagnosis [5].

II. Material and Methods

Heart disease is referring as the diseases that cause by the functional failure of heart which also refer as heart attack and includes coronary artery disease, arrhythmias, atrial fibrillation, heart valve disease, congenital heart disease, cardiomegaly (enlarged heart), cardiomyopathy (heart muscle disease), etc. [6]. In the United States, heart disease is the top one disease in death rate either for male or female. Symptoms of heart disease can be subtle and go unnoticed until a major event like a heart attack occurs and finally lead to death. Noticeable symptoms of common heart disease are extreme fatigue, chest pain, and difficulty in breathing. Unhealthy lifestyle habits such as imbalance nutrition and diet could lead to heart disease. While certain uncontrollable risk factors such as age and gender might be one of the factors of heart disease, in order to keep the heart healthy, it is a good idea to lower the blood pressure, obtain high-fiber nutrition, workout once a week, manage the stress, and stop smoking.

A. Data Mining

For decades, researchers have studied the pattern of heart disease by implement data mining in the healthcare industry. Due to the abundance of data and the robustness of data analytic tools, one situation namely "data rich but poor information" is happened. The large dataset become data tombs, while the analyze data could neither help in disease diagnosis nor treatment. In the healthcare industry, data mining has been improved by different researchers and namely as pattern analysis, knowledge mining from databases, knowledge extraction, data dredging, and data archaeology.

B. Classifier Algorithm

There are many algorithms for the classification of heart disease datasets, such as Naïve Bayes, neural networks, SMO, decision trees, and many more. In this research, five Weka classifiers have been evaluated under secondary heart dataset. The brief descriptions of each of the classifiers used are given below.

1) Decision Stump: A decision stump is one of the machine learning models which apply decision tree on it [7] — decision stump connected between terminal nodes and internal nodes, where terminal nodes represent leaves while internal nodes represent root. The value of input attribute is used to predict in decision stump, which named as 1-rule in Decision Stump.

2) Random Forest: Random Forests are broadly believed to be the finest "off-the-shelf" classifiers envisaging high dimensional data [8]. It is an assortment of tree predictors such that distributed equally for all trees in the forest each tree relies on the values of a random vector sampled autonomously. Training data is differentiating to be selected, with replacement, to train each tree. While the remaining training data are then used to determine the variable of importance and errors, the class assignment is made by the number of votes from all the trees and for deteriorating the average of the results is used. It is similar to bagged decision trees with barely some difference.

3) Neural Network: Neural Network (NN), is named after the mimic of neurological functions in the human brain (i.e., neural networks) [9]. Computational nodes in NN are used to emulate the functions of the neurons in the brain. The nodes acted as a processor and interconnected with another node. The nodes can tune when the NN is learning from the data pattern. The nodes are classified based on its layer, which is the inner and outer layer.

4) Multilayer Perception: Multilayer Perceptron is a nonlinear classifier which is back propagation neural network with one or more layers between input and output layer [10]. Commonly, perceptron network is illustrated as three layers. Artificial Neural Networks (ANN) is one of the examples of multilayer perception which also well known as classification algorithm. In order to employ ANN, Multi-Layer Perceptron (MLP) were used in this work.

5) SMO: SMO is an improved algorithm by customized training on Support Vector Machines (SVMs). In order to train a support vector machine, a solution is needed to solve the very large quadratic programming optimization (QP) problem. SMO is applied to break the large QP problem into a sequence of smallest possible QP problems, which these small QP problems are solved analytically. The memory of SMO is vary based on the training dataset, which allows SMO to deal with a large amount of training dataset.

6) J48: J48 is an implementation of C4.5 in WEKA. C4.5 uses information entropy concept. In WEKA, J48 algorithm is implemented as C4.5 decision tree learner. This algorithm is using a greedy technique to combine decision trees and applied reduced-error pruning. J48 is used to build a decision tree from a set of labeled training data using the information entropy. J48 is going to split the attributes and build a decision tree and calculate the normalized information gained. When the subset is classed, the splitting process is considered completed. A leaf node is present or being created to choose that class a possibility also can be there that none of the features provides information gain. J48 can apply in discrete and continuous attributes. It can be used in attributes with differencing lost as well as the missing attribute value.

C. Attribute Selection

Attribute selection is used to reduce the dimensionally of the training and the test data before being passed on to the classifier. There are many attribute evaluation methods in the

Waikato Environment for Knowledge Analysis (WEKA) tools, which 5 evaluators are selected for this research. There are Gain Ratio Attribute Evaluator, Correlation Attribute Evaluator, OneR Attribute Evaluator, Cfs Subset Evaluator, and Principal Components.

1) GainRatioSubsetEval: Gain Ratio Attribute Evaluator evaluates the worth of an attribute by measuring the gain ratio concerning the class [11]. Information gain (relative entropy, or Kullback-Leibler divergence), in information and probability theory, is a measure of the difference between two probability distributions.

2) CorrectionAttributeEval: Correction Attribute Eval is used to evaluate the worth of an attribute by measuring the correlation between attribute and the class [12]. The values are used to treat as an indicator to obtain a significant value for nominal attributes. Hence, the nominal attribute is proportional average.

3) OneRAttributeEval: The other selected evaluator used in this research is One R. OneR classifier is used to investigate the usage of attributes [13].

4) CfsSubsetEvaluator: CfsSubsetEvaluator is a filter algorithm which ranks the attribute based on the correlation within heuristic evaluation function [14]. The significant attribute is then used in individual predictive as well as checking the degree of redundancy between each other. The significant subsets are those that are highly correlated with the class while having low intercorrelation.

TABLE I
SELECTED DATASET FROM THE HUNGARIAN INSTITUTE OF CARDIOLOGY

Sr.no	Attribute	Description	Values
1	age	Age in years	Continuous
2	sex	Male or female	Discrete: 1 = male 0 = female
3	chest_pain	Chest pain type	Discrete: 1 = typical type 1 2 = typical type agina 3 = non-agina pain 4 = asymtomatic
4	trestbps	Resting blood pressure	Continuous value in mm hg
5	chol	Serum cholesterol	Continuous value in mm/dl
6	fbs	Fasting blood sugar	1 ≥ 120 mg/dl 0 ≤ 120 mg/dl
7	restecg	Resting electrocardiograp hic results	0 = normal 1 = having ST T wave abnormality 2 = left ventricular hypertrophy
8	thalach	Maximum heart rate achieved	Continuous value
9	exang	Exercise induced angina	0 = no 1 = yes
10	oldpeak	ST depression induced by exercise relative to rest	Continuous value
11	slope	Slope of the peak exercise ST segment	1 = unsloping 2 = flat 3 = down sloping
12	ca	Number of major vessels colored by floursopy	0 – 3 value
13	thal	Defect type	3 = normal 6 = fixed 7 = reversible defect

5) Principal Components: Principal Components performs a principal components analysis and transformation of the data — the evaluator used in conjunction with a Ranker search. Dimensionality reduction is accomplished by choosing enough eigenvectors to account for some percentage of the variance in the original data with a default of 0.95 (95%). Attribute noise can be filtered by transforming to the PC space, eliminating some of the worst eigenvectors, and then transforming back to the original space.

The pre-processing technique is applied to clean up the data and followed by classification techniques to classify the dataset. Figure 1 gives the illustration of the computational framework which is associated with all methods in this study. The proposed framework contains five steps namely get the medical database, data preprocessing, classification, accuracy comparisons, results in determination and research conclusion.

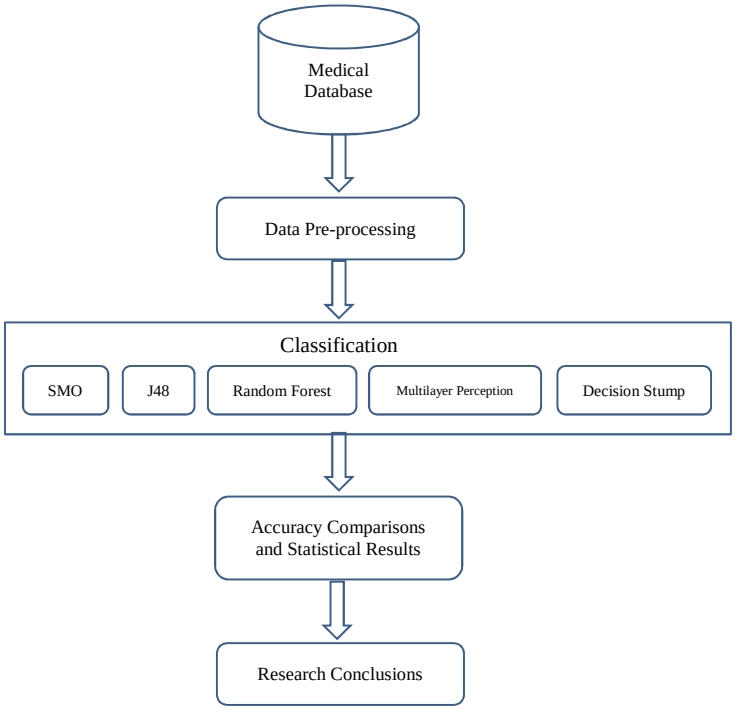

Fig. 1 Computational Framework

D. Heart Disease Classification

Weka (version 3.8.1) is used as a mining tool for this classifying data analysis method. Heart Disease data sets are applied to study the performance of a comprehensive set of classification algorithms (classifiers). In this paper, we have four Weka classifiers under secondary dataset which were J48, SMO, Decision Stump, Random Forest, and Multilayer

Perception. The brief description of each classifier is given in this section.

E. Attribute Selection Methods

The main aim of feature selection techniques is to remove irrelevant or redundant features from the dataset. Attribute selection is one of the processes to remove the redundant attributes that are irrelevant to the given data mining task. There are two categories of feature selection: wrapper and filter.

III. Results and Discussion

The Hungarian Institute of Cardiology collects the heart disease dataset revealed that there is a total number of 76 raw attributes in this dataset. However, in this study, we are only using 11 attributes. In the meanwhile, this dataset contains 294 rows, which means it contains 294 patient's information. Table 1 determine the details of the dataset and its description. The selected dataset from the Hungarian Institute of Cardiology is shown below:

F. Arff Creation

The Attribute-Relation File Format also named as ARFF file is one of the file format supported in Weka. It is an ASCII text file which describes a list of instances sharing a set of attributes. The researcher in the Department of Computer Science of The University of Waikato had developed ARFF files and implemented it into Weka machine learning software.

G. Research Results and analysis

After the experimental work, we have observed that the classification accuracy of Multilayer Perception is higher than the other four classifiers. The accuracy of each classifier is shown in Table 2 and figure 2. Table 3 and figure 3 shows the mean absolute error of each classifier. We found that the mean absolute error of Multilayer Perception classifier is less compared to the other four classifiers

TABLE II
ACCURACY OF CORRECTLY CLASSIFICATION (%)

	DATASETS				
Evaluator/ classifier	Decision Stump	Random Forest	Multilayer Perceptron	SMO	J48
No attribution selection :	79.93	79.93	85.03	82.65	80.95
With attribute selection :					
Gain Ratio Attribute Eval	77.89	79.59	78.91	80.95	77.55
Correlation Attribute Eval	76.87	81.29	79.59	79.59	76.53
OneR Attribute Eval	79.93	77.89	80.27	81.97	79.93
Cfs Subset Eval	79.93	78.91	80.95	81.97	77.89
Principal Components	82.65	82.99	82.99	82.65	81.97

TABLE III
MEAN ABSOLUTE ERROR OF TESTED CLASSIFIER

	DATASETS				
Evaluator / Classifier	Decision Stump	Random Forest	Multilayer Perceptron	SMO	J48
No attribution selection :	0.12	0.12	0.07	0.18	0.12
With attribute selection :					
Gain Ratio Attribute Eval	0.13	0.09	0.09	0.18	0.12
Correlation Attribute Eval	0.13	0.09	0.09	0.18	0.12
OneR Attribute Eval	0.12	0.10	0.09	0.18	0.12
Cfs Subset Eval	0.12	0.09	0.09	0.18	0.12
Principal Components	0.11	0.09	0.09	0.18	0.11

TABLE IV
KAPPA STATISTICS OF TESTED CLASSIFIER

	DATASETS				
Evaluator / Classifier	Decision Stump	Random Forest	Multilayer Perceptron	SMO	J48
No attribution selection :	0.56	0.56	0.67	0.61	0.57
With attribute selection :					
Gain Ratio Attribute Eval	0.51	0.56	0.53	0.57	0.49
Correlation Attribute Eval	0.49	0.59	0.55	0.54	0.46
OneR Attribute Eval	0.55	0.52	0.56	0.59	0.55
Cfs Subset Eval	0.55	0.53	0.58	0.59	0.50
Principal Components	0.62	0.62	0.62	0.62	0.61

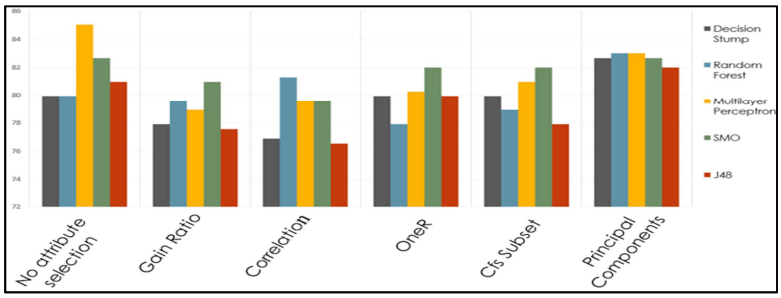

Fig. 2 Graphical representation of Classification accuracy of the tested classifier

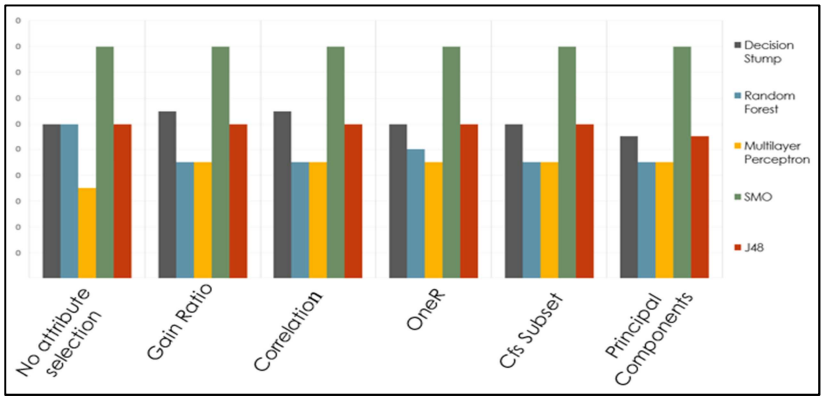

Fig. 3 Graphical representation of Mean absolute error of *the tested* classifier

In this study, five different classification algorithms under WEKA compared with using UCI data set. Then pre-processed datasets used to test the five classifiers using 10-folds cross-validation — their different performance measures considered for classifiers. Results of comparison showed that Multilayer Perceptron classifier achieved the highest value in accuracy and lowest value of mean absolute error measures which we can see that in table 1 and 2.

Five different classifiers had been used, and the algorithm is being compared and select the best classifier whereas two different scenarios are conducted. Firstly, all the attributes were applied to it. While for the second scenario, specific attributes were selected for it. In order to compare these experiments, performance measurement such as accuracy, mean absolute error, and kappa statistics of the tested classifier are used. The summary of the implemented algorithms is already shown in the above tables and figures.

IV. Conclusion

In this study, heart disease dataset is used and analyzed to obtain data behavior. The overall objective of this study is to obtain a better prediction method for heart disease. The attributes are shortlisted to get better accuracy results. Throughout the study, 5 classification method is applied, which are SMO, J48, Multilayer Perception, Decision Stump and Random Forest. The experiment runs during the study were separated into 2, which applied all attributes and certain attributes only. The file type for the dataset is ARFF which supported by Weka. Results show Multilayer Perception Neural Network has better accuracy compared to other classification methods. The outcome results could be used as references to help in improving the diagnosis of heart disease. Multilayer Perception Neural Network can be further improved to obtain higher accuracy. The researcher can focus on the number of significant attributes that are going to apply to them. In the meanwhile, other classification techniques such as Decisions Tree, Naïve Bayes, and Significant Directed Random Walk can be applied to predict and classify heart disease.

Acknowledgment

Universiti Teknologi Malaysia sponsors this work.

References

[1] Dangare, C. S., & Apte, S. S. (2012). Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), 44-48.

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

[3] Seah, C. S., Kasim, S., Fudzee, M. F., Ping, J. M., Mohamad, M. S., Saedudin, R. R., & Ismail, M. A. (2017). An enhanced topologically significant directed random walk in cancer classification using gene expression datasets. Saudi Journal of Biological Sciences, 24(8), 1828-1841.

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

[5] Palaniappan, S., Awang, R., Intelligent Disease Prediction System Using Data Mining Techniques, IJCSNS International Journal of Computer Science and Network Security. 8(8): 343-350 (2008).

[6] Chaurasia, V., & Pal, S. (2014). Data mining approach to detect heart diseases.

[7] Symbology of the Logical Decision Tree. (2017). Decision-Making Management, 99-100. doi:10.1016/b978-0-12-811540-4.09979-8 Available from: http://en.wikipedia.org.[Last accessed on May 11].

[8] Leo Breiman (2001). Random Forests. Machine Learning. 45(1), pp.5-32.

[9] Palaniappan, S., Awang, R., Intelligent Disease Prediction System Using Data Mining Techniques, IJCSNS International Journal of Computer Science and Network Security. 8(8): 343-350 (2008).

[10] Capilla, C. (2014). Multilayer perceptron and regression modelling to forecast hourly nitrogen dioxide concentrations. Air Pollution XXII. doi:10.2495/air140041

[11] GainRatioAttributeEval. (2017, December 22). Retrieved from http://weka.sourceforge.net/doc.dev/weka/attributeSelection/GainRatioAttributeEval.html

[12] CorrelationAttributeEval. (2017, December 22). Retrieved from http://weka.sourceforge.net/doc.dev/weka/attributeSelection/CorrelationAttributeEval.html

[13] OneRAttributeEval. (2017, December 22). Retrieved from http:// http://weka.sourceforge.net/doc.stable-3-8/index.html?weka/attributeSelection/OneRAttributeEval.html

[14] CfsSubsetEval. (2017, December 22). Retrieved from http://weka.sourceforge.net/doc.dev/weka/attributeSelection/CfsSubsetEval.html