Diagnostic Prediction Models for Cardiovascular Disease Risk using Data Mining Techniques

Nongyao Nai-arun[1] and **Rungruttikarn Moungmai**[2]

ABSTRACT

Cardiovascular disease is the top national health problem that leads to a large number of deaths in Thailand. There is still a growing number of patients with the disease. Proactive measures of disease prevention and disease control are searching for risk groups. Therefore, people who are at risk can diagnose and manage themselves to reduce risk factors and adjust their behavior accordingly. For this reason, the idea of diagnostic prediction models for cardiovascular disease was conceived. The data of patients from 126 health promoting hospitals and 12 hospitals in Saraburi Province was collected. Then analysis was done to establish 6 models: logistic regression, random forest, back-propagation neural network, decision tree, naïve-bayes and K-nearest neighbors. Moreover, 10-fold cross validation was applied to each model. The results revealed that the logistic regression model achieved the highest accuracy rate, 99.940%, followed by the back-propagation neural network model, 98.105%.

Keywords: Cardiovascular Disease, Logistic Regression, Random Forest, Back-propagation Neural Network, Naïve-bayes, K-nearest Neighbors, Decision Tree

1. INTRODUCTION

Current mortality rate statistics [1] show that the number of deaths from non-communicable disease (NCDs) is approximately 41 million per year or 71% of all worldwide deaths. It also indicates that 85% of these deaths, for 15 million people, are premature for people whose ages are between 30 and 69 years old and came from low and medium income countries. In addition, the statistics present the number of deaths caused by various diseases of the NCDs group. First, the number of deaths from Cardiovascular disease is 17.9 million people. Second, about 9 million people die from cancer. Third, respiratory disease killed approximately 3.9 million people. Finally, 1.6 million people died from Diabetes Mellitus.

In addition to the above numbers, the World Health Organization (WHO) [2] reveals that about 31% of global deaths, or 17.9 million people, die from cardiovascular disease every year. The stimulation of the disease comes from the bad behaviour of each person, such as unhealthy food, lack of exercise, smoking, and drinking alcohol. Basic symptoms of heart attack and stroke, which cause the deaths of the disease, can be observed or diagnosed from abnormal values of heart health such as high blood pressure, high blood glucose, and being overweight.

The Public Health Statistics Report of the Ministry of Public Health in Thailand [3] states that the mortality rate of coronary artery disease per 100,000 population was 23.4, 26.9, 27.8, 29.9, and 32.3 between 2012 and 2016 respectively. It also states that the cardiovascular disease per 100,000 population was 412.70, 427.53, 431.91, 407.70, and 501.13 in 2011 to 2015 respectively. It can be seen from these numbers that the intensity of the coronary heart disease keeps increasing due to a trend where both the mortality rate and the number of patients are steadily increasing. Based on the literature review, the current situation, and the service of chronic non-communicable diseases, the department of medical services found that the average cost of treatment for heart disease patients is approximately 6,906 million baht per year. Moreover, the disease is the top cause of health loss of the working-age population. It also affects the quality of life of the population and causes economic loss. For these reasons, the disease affects personal life, family, society, and the nation [4].

A vast amount of data is collected in hospitals. Many researchers are interested in using this information for their benefit and are applying data mining techniques to the huge data sets. These techniques are used for analysing the data and searching for patterns or rules. They are also finding ways of retrieving data or searching for interesting and useful knowledge from a large database [5], [6], [7]. Data mining techniques relate to other various disciplines including database systems, data warehouse technology, statistical analysis, machine learning, pattern recognition, artificial neural networks, data retrieval, image processing and signals, analysis of spatial data of events,

Manuscript received on June 30, 2019 ; revised on March 3, 2020.

Final manuscript received on March 31, 2020.

[1] The author is with Faculty of Science and Technology, Phranakhon Si Ayutthaya Rajabhat University, Thailand. E-mail: nongyaonaiarun@gmail.com

[2] The author is with Faculty of Science and Technology, Nakhon Sawan Rajabhat University, Thailand. E-mail: r.moungmai@gmail.com

DOI: 10.37936/ecti-cit.2020142.199897

and creating effective forecasting models. Data mining is based on outstanding analytical techniques and is more complex than statistical analysis and general structured queries (SQL language). There are many popular data mining techniques such as data classification, clustering, association rules, and sequence analysis. [8], [9].

The researchers therefore had the idea of using data mining techniques to create predictive models to develop a system for predicting cardiovascular disease risk. This can be used in hospitals to screen the disease risks of new patients since they have no symptoms. Hence, the cardiovascular patients will be helped to heal themselves from the initial stage. In addition, both patients and non patients can use it for diagnosis and health promotion to prevent themselves suffering from cardiovascular disease.

2. LITERATURE REVIEWS

Data mining is widely used in various fields including in medical and healthcare areas [10]. It is very important because it is related to human life including the well-being and economy of the country. In addition, it is not easy to handle the medical data because it is in very huge databases and keeps increasing continuously. Therefore there are various researchers who have studied how to deal with a vast database and they have tried to establish forecasting models of people's illness [11]. Here are some reviews of such studies.

Suksawatchon et. al [12] introduced two new approaches, Health Risk Analysis System (HRAS) and Risk Analysis Classifier (RAC), to identify health risk levels in 3 health aspects, namely mental, physical, and social health aspects. They also combined the RAC model with other classifiers that are decision tree, naïve-bayes, neural network, LibSVM, and ensemble with voting approaches. The results state that the best model is a neural network algorithm. Its accuracy rates are more than 90% in all aspects of prediction.

Rachata et. al [13] studied a prediction model to identify cardiovascular patients who have type 2 Diabetes Mellitus and hypertension. A Fuzzy logic based method was used to construct a proper model. 15 input variables of both clinical and lifestyle risk factors were used as predictors. Also, the scientific data and the experts' implicit knowledge were included in this study. The proposed model achieved very high accuracy rate of 96.69% which is derived from comparison with the experts' decisions.

Saxena and Sharma [14] suggested an Efficient Heart Disease Prediction System that is a system for prediction of heart disease. 13 factors related to heart disease were analysed as input variables. Moreover, many well known classifiers were conducted including Support Vector Machines (SVM), decision tree, PART, back-propagation neural network, random forest, and TSEAFS. Also, to create a prediction model, a 10-fold method was used. After each model was created, its accuracy rate was calculated and compared with the proposed model. They found that the accuracy rate of the system model, 86.7%, was greater than other approaches.

Assari et. al [15] created a model for heart disease diagnosis. Their input variables include 5 continuous and 8 discrete factors related to the disease. The output is binary data of heart disease diagnosis. Values of the binary output consist of healthy people and patients who are subject to possible heart disease. The proposed models consist of naïve-bayes, K-nearest neighbours where K = 7, Support Vector Machines (SVM), and decision tree. They found that the highest accuracy rate was 84.33% from the SVM model.

Nai-arun and Moungmai [16] applied data mining techniques for calculating the risk of diabetes. Both classification and ensemble algorithms were conducted to predict a dichotomous output. The output variable is diabetes risk where the sample population was divided into two groups: normal and diabetes risk groups. Their input variables consist of 6 numeric and 5 qualitative values. These data values are from basic screening information without medical diagnosis. They found that the best performance of their data set was with the random forest model with an accuracy rate of 85.558%, followed by bagging with decision tree and bagging with artificial neural networks, yielding 85.333% and 85.324% respectively. It can be seen that popular classifiers are decision tree, neural networks, naïve-bayes, and K-nearest neighbours. Therefore these algorithms will be applied in our study.

3. BACKGROUND KNOWLEDGE

3.1 Data Mining

Data mining is a technique to manipulate a large database or big data for searching for a suitable pattern in the data. This technique is applied to various fields such as medicine, economy, agriculture and education for the analysis of data and forecasts of interesting results. [17] There are many techniques such as artificial neural networks, logistic regression, decision tree, K-nearest neighbours, and naïve-bayes. Here only two algorithms used in this study are presented. The rest can be seen in more detail in Nai-arun and Moungmai [16].

3.2 Artificial Neural Network

An Artificial Neural Network (ANN) is a model that mimics the human nervous system and is used for qualitative data forecasting [18]. The network is divided into layers in which each class consists of at least 1 node, and each node has a link to other nodes of different layers. These links will be randomly given a value of weight and are called neuron or weight lines. In general, the network consists of 3 layers: input layer, hidden layer, and output layer, as shown in Fig.1. Input layer can be independent variables or factors that effect the dependent variable in the output layer. For example, this study would like to predict the risk of cardiovascular disease. Therefore the input layer consists of basic factors that are related to the disease, such as age, blood pressure, and cholesterol level. The output layer is the disease risk. A hidden layer exists between the input and output layers, and it might be divided into many layers depending on the complexity of data analysis and its network.

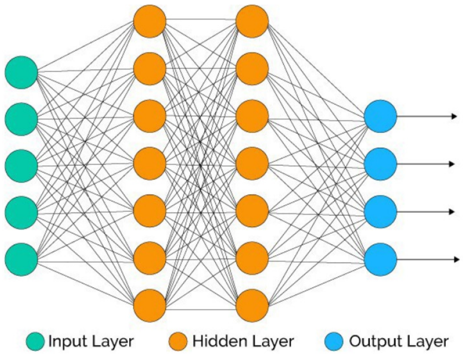

Fig.1: *A Process of Artificial Neural Network [19].*

There are many types of neural network models such as single-layer perceptron neural network, multilayer perceptron neural network, associative neural network, and hopfield network. In this study, a multilayer perceptron, or back propagation neural network, was applied.

A multi-layer perceptron neural network, also known as a back-propagation neural network (BPNN), is a network with both reverse and forward learning. All input variables have to be related to the output variable. In addition, the networkmodifies itself using random weights for all layers in the network. Although this method is suitable for complex data, it is not good for many artificial neural networks working together.

Let N be the number of nodes in input layer

M be the number of nodes in hidden layer

J be the number of nodes in output layer

x_n is the input value of node n in input layer where $n = 1, 2, \ldots, N$

s_m is the output value of node m in hidden layer

y_m is the modified output value of node m in hidden layer

v_j is the output value of node j in output layer

z_j is the modified output value of node j in output layer

w_{nm} is the weight of weight line between input node n and hidden node m

w_{mj} is the weight of weight line between hidden node m and output node j

$f(x)$ is the sigmoid transfer or activation function, and can be calculated from

$$f(x) = \frac{1}{1 - e^{-x}} \tag{1}$$

To establish a model, there are 7 steps as follows:

Step 1: The number of input nodes, the number of output nodes, the number of sub-layers and hidden nodes of each layer, the number of repeat cycles and error are defined.

Step 2: The weight of each weight line is randomly defined between -1 and 1.

Step 3: The output value, s_m, of hidden layer is calculated from

$$s_m = \sum\nolimits_{n=1}^{N} x_n \times w_{nm} \tag{2}$$

Step 4: The output value, s_m, is modified by using Sigmoid transfer function

$$y_m = f\left(\frac{1}{1 - e^{-s_m}}.\right) \tag{3}$$

Step 5 : The output value, v_j, of output layer is calculated from

$$v_j = \sum\nolimits_{m=1}^{M} y_m \times w_{mj} \tag{4}$$

Step 6: The output value, v_j, is modified by using Sigmoid transfer function

$$z_j = f\left(\frac{1}{1 + e^{-v_j}}.\right) \tag{5}$$

Step 7: All modified output values of the output layer are compared with the actual output values and are used to calculate an error. If this error is more than a criterion, a new set of data will be calculated by repeating Steps 2 to 7. This iteration will end when the error is the low enough.

3.3 Logistic Regression

Logistic regression is a forecasting technique in which the response variables are qualitative or category data. Its input variables, or predictors, can be both qualitative and quantitative data and can have one or more variables. In general, there are 2 types of models depending on the number of values of the dependent variable. One is called binary logistic regression; its dependent variable has 2 values. Another is called multinomial logistic regression; its dependent variable has more than 2 values [20].

Let k be the number of input variables

m be the number of classes of the output variable

X_i be input variables where $i = 1, 2, \ldots, k$

Y be the output variable

P_j be probability of class j of the output variable where $j = 1, 2, \ldots, m$

$g(x)$ is the link function.

To construct the logistic regression model, the link function will be first calculated from

$$g(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{6}$$

where $\beta_i, i = 0, 1, \ldots, k$, are coefficients of the model that will be estimated by maximum likelihood method [20].

Then, forecasting models of each class of the output variable will be calculated from

$$\ln\left(\frac{P_j}{1 - P_j} = g(x)\right) \tag{7}$$

In this study, the output variable is divided into m classes, hence there are m–1 models.

3.4 Cardiovascular Disease

Cardiovascular disease is a disease caused by atherosclerosis or blood clotting, which is caused by the accumulation of fat and minerals in the vascular wall until it narrows. For this reason, there is resistance to blood flow, a lack of vascular flexibility, and blood vessels are more fragile. If these symptoms occurr in the arteries that feed the heart, they will cause less blood to the heart, leading to ischemic heart disease. If the artery is clogged until the blood does not flow to the heart, it will cause an acute heart attack or heart failure and myocardial infarction. Moreover, if the occurrence is in blood vessels to the brain, it will cause less blood to reach the brain and cause blood deficiency or paralysis.

The symptoms of cardiovascular disease usually do not appear in the initial stage, and the diagnosis does not happen until there are severe symptoms such as unconscious patients. The disease often is only detected when an acute event occurs in which the patients often die. In addition, the cost of treatment for this disease is very high. In 2019, the Thai Ministry of Public Health has a policy for every hospital to screen patients by collecting basic information including gender, smoking behaviour, Diabetes Mellitus history, age, blood pressure, cholesterol, waist size, and height. This is done to help the patients to be diagnosed, since early diagnosis reduces the mortality rate and medical expenses [21].

4. MATERIAL AND METHOD

The process of our experiments is shown in Fig.2.

4.1 Data Preprocessing

An initial real data set was collected from 126 health promoting hospitals (HPH) and 12 other hospitals in Saraburi Province during 2018 – 2019. Each person filled out a screening form used to identify cardiovascular risk groups in the study. In this process, the data set was manipulated as described in the following paragraphs of this section.

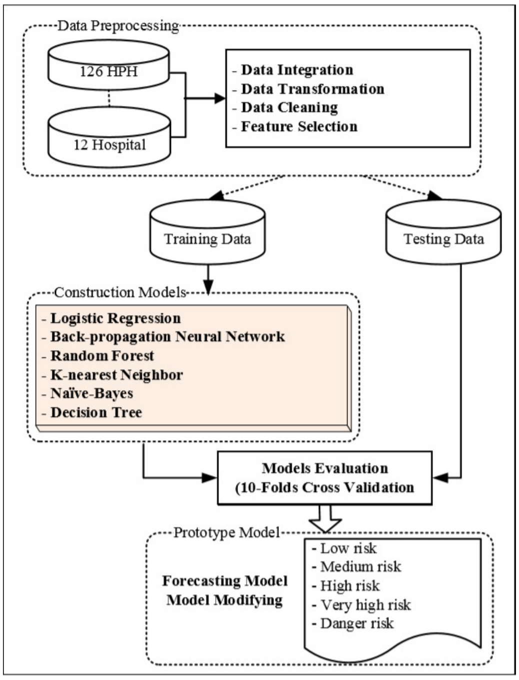

Fig.2: *The Conceptual Framework of this study.*

4.1.1 Dataset

In 2017, the Division of Non Communicable Diseases, Department of Disease Control, Thai Ministry of Public Health, [3] prepared a risk assessment form for cardiovascular disease and required hospitals throughout the country to use it in screening for the disease risk. Hence, a dataset used in this study was based on data from the forms and was selected from all hospitals in Saraburi province by using purposive sampling of responses. The dataset consists of data from 44,674 patients collected between 2018 and 2019 from 138 hospitals. According to the risk form, there are 8 factors of basic information related to cardiovascular disease risk including gender, smoking behaviour, diabetes mellitus, age, height, waist size, blood pressure, and cholesterol level. Therefore, only these factors were used in modelling, and their details are shown in Table 1. These factors can be classified into 2 types of data: qualitative and quantitative data. The qualitative data consists of gender, smoking behaviour, and diabetes mellitus. The quantitative data consists of age, height, waist size, blood pressure, and cholesterol level. All factors of were used as input variables in modelling, and the output variable was class, which is cardiovascular disease risk. Classes were used for dividing people into 5 groups [22].

Table 1: *Input and Output Variables .*

No	Attributes	Description	Values
1.	Sex	Gender	Nominal scale M: Male F: Female
2.	Smoking	Smoking behaviour	Nominal scale Y: Yes smoke N: No smoke
3.	Dm	Diabetes Mellitus	Nominal scale Y: Yes N: No
4.	Age	Age (year)	Numerical scale
5.	Height	Height (cm)	Numerical scale
6.	Waistcm	Waist_cm (cm)	Numerical scale
7.	Sbp	Systolic blood pressure (mmHg)	Numerical scale
8.	Chol	Total cholesterol (mg/dL)	Numerical scale
9.	Class	Cardiovascular risk	Ordinal scale R1: Row risk R2: Medium risk R3: High risk R4: Very high risk R5: Danger risk

4.1.2 Data Preparation

In this stage, the dataset was manipulated as follows:

Step 1 : Data Integration

Data was collected from 138 hospitals using different databases. All of the data needed to be combined in one file.

Step 2 : Data Transformation

There was some information that could not be put into the forecasting model directly. Hence, it needed to be transformed. For example, date of birth of each patient was changed to age (in years). Moreover, all qualitative input variables needed to be transformed as a dummy variables before putting then into the models.

Step 3 : Data Cleaning

Before analysing the dataset, all information of each patient must be complete. Therefore, all incomplete records that had missing values and no information for one or more attributes were excluded from the dataset. For example, a patient whose information about sex or height disappeared was deleted from the study.

Step 4 : Feature Selection

It is very important to select good attributes for a forecasting model. So the filter approach was applied and WEKA was run in this stage. Pearson correlation of each attribute and the output variable was calculated and ranked. Then each attribute was chosen to be put into the model in order of the relation values from the highest to the lowest.

4.1.3 Cardiovascular Risk

In this study, cardiovascular disease risk calculated as a numeric value is the dependent variable and will be predicted using data mining techniques. Therefore the risk values are divided into five classes based on the criterion in the cardiovascular disease handbook of the Thai Ministry of Public Health [21], [22]. Let R1, R2, R3, R4, and R5 be the risk classes that are under an ordinal scale. The detail of each class is as follows:

- R1 means a low risk where its rate is less than 10
- R2 means a medium risk where its rate is between 10 to 20
- R3 means a high risk where its rate is between 20 to 30,
- R4 means a very high risk where its rate is between 30 to 40
- R5 means a danger risk where its rate is more than 40 and the patient needs to see a doctor.

4.2 Construction of Models

Here, 6 popular approaches were applied: logistic regression, back-propagation neural network, K-nearest neighbours', naïve-bayes, decision tree, and random forest. To evaluate these models, the dataset was first divided into 2 groups. One was 80% of the dataset for creating a forecasting model, called training data. Another was for precision checking of the model, called testing data. Then, the training data was used to establish a proper model which was evaluated by using the testing data. All input values of the testing data were used as input to the model for predicting output values. After that, these values were compared with the actual output values of the testing data. The assessment concluded by examining the validity of the model in forecasting. There are many validations such as Accuracy rate, Mean Square Error (MSE) and ROC curve.

In this study, 10-fold cross validation was applied because all data was used as both training and testing data. A process of 10-fold cross validation starts with dividing all data into 10 groups. 9 groups are used for training and another is used for testing. First, groups 1 to 9 are used as training data and group 10 is used as testing data. Second, groups 1 to 8 and 10 are used as training data and group 9 has been used as testing data. Then, the process is repeated until every group is used as testing data. Also, an accuracy rate is estimated after every iteration. It can be seen that there are 10 accuracy rates. Therefore the accuracy rate of the model will use an average value of the 10 rates.

4.3 Prototype Model

The dataset used here was only from patients in Saraburi province between 2018 and 2019. Therefore the best model obtained in this study is only a prototype forecasting model. It can be used for preliminary predictions now, but before wider application it should have further development using techniques such as applying ensemble approaches and adding more datasets and other attributes that are related to the disease risk.

5. RESULTS

5.1 General Information

First of all, the data set of 44,674 patients was put in the data preprocessing stage. 12,745 incomplete records were taken out of the study so there were 31,929 patients left in this study, 21,362 female (66.90%) and 10,567 male (33.10%). It was found that there were many people with Diabetes Mellitus 39.58% and smoking behaviour 4.90%. The average and standard deviation of all quantitative variables were calculated as shown in Table 2. These average values indicate that most people in the sample were old. The average age was 63.30 years old. This also shows that most patients are overweight since their average waist is around 85.89 centimetres (33 inches), while the mean of height is around 158 centimetres. Moreover, the average cholesterol level is quite high, about 187.45 mg/dL. Although the average systolic blood pressure, 129.35 mmHg, is in the normal range, it is quite high.

Table 2: *Minimum, Maximum, Mean and Standard Deviation (SD) of Each Attribute.*

Attributes	Min	Max	Mean	SD
Age (year)	35	102	63.30	11.39
Height (cm)	145	196	158.86	7.46
Waist (cm)	50	127	85.89	10.59
Systolic blood pressure (mmHg)	80	216	129.35	14.60
Total cholesterol (mg/dL)	70	320	187.45	40.51

Next, the cardiovascular risk, which is numeric, was classified into 5 classes and each class count of the number of patients is shown in Table 3. The biggest risk class is low risk (R1), 47.10%. The rest can be arranged in the sequence medium risk (R2), high risk (R3), danger risk (R5), and very high risk (R4) respectively.

Before the model of each approach was established, all attributes were studied to discover how much they related to the class using Pearson correlation. These correlations were ranked as shown in Fig 3. It can be seen that age, systolic blood pressure (SBP), and Diabetes Mellitus (DM) are the first three ranks respectively, while the final rank is height. Therefore, age was first put into a model and the remaining attributes were put into the model one by one respectively.

5.2 Model

To seek the best, most suitable model for predicting cardiovascular risk, 6 algorithms were tried. Models of back-propagation neural network and decision tree are shown in Fig.4 and Fig.5.

Table 3: *Number and Percentages of Each Class.*

Class	Number (Percentages)
R1	15,040 (47.10%)
R2	8,260 (25.87%)
R3	4,056 (12.70%)
R4	2,067 (6.48%)
R5	2,506 (7.85%)
Total	**31,929 (100%)**

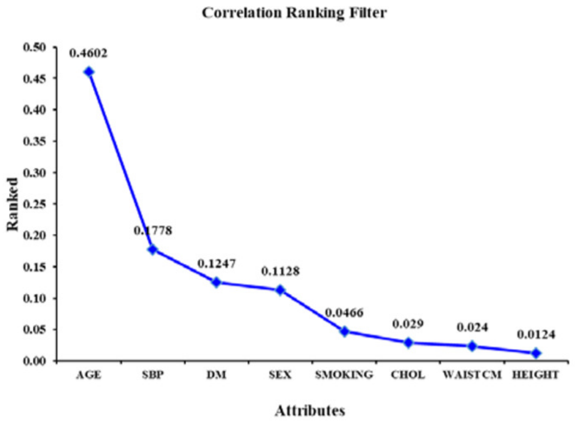

Fig.3: *Feature Selection.*

Fig.4 shows the back-propagation neural network model that includes 8 input nodes and 5 output nodes for the risk classes. The hidden layer and its nodes are defined randomly in the middle between the input and output nodes.

Fig.5 presents a part of the decision tree model that looks like a big tree. The network begins with classification of each input variable until the leaf nodes at the end of each branch will be predictors of risk class.

5.3 Performance Evaluation

In addition, given model accuracy rates are shown in Table 4. As stated in the schedule, the forecasting approach with the best performance is the logistic regression model with an accuracy of 99.940%, followed by backpropagation neural network, random forest, decision tree, and K-nearest neighbors respectively. The lowest accuracy was from naïve-bayes.

In order to confirm the selection of the best model, Mean Square Error (MSE) and ROC curve of proposed models were calculated, and they are shown in Fig.6. This still indicates that the minimum value of MSE and ROC are from the logistic regression

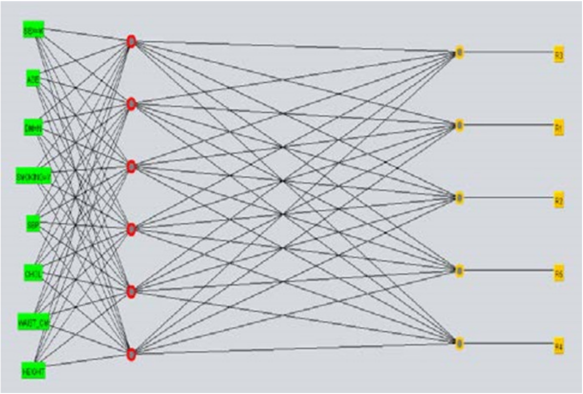

Fig.4: *Model of Back-propagation Neural Network.*

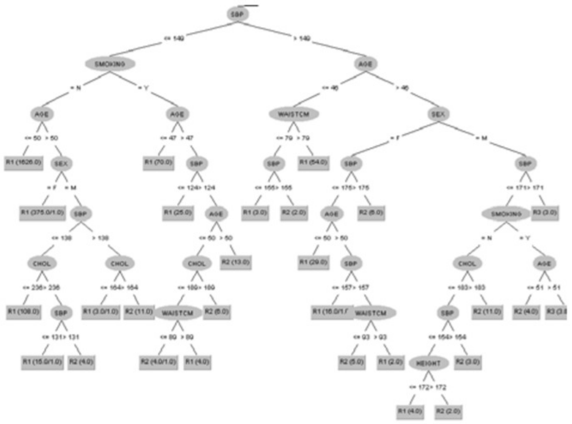

Fig.5: *Model of Decision Tree.*

Table 4: *Models Accuracies.*

Models	Classified Instances		Accuracy
	Correct	Incorrect	(%)
Logistic Regression	31910	19	99.940
Back-propagation Neural Network	31324	605	98.105
Random Forest	30533	1396	95.627
Decision Tree	30068	1861	94.171
K-nearest Neighbors	27500	4429	86.128
Naïve-bayes	22859	9070	71.593

model, followed by the back-propagation neural network model.

5.4 Confusion Matrix

Table 5, 6, and 7 display confusion matrixes for the top three highest accuracy rates. As stated in the matrix of logistic regression, there are very few invalid predictions, only one to four mistakes for each class. Incorrect predictions of the random forest, back-propagation neural network models occurred more often than in the logistic regression model, but still are not considered frequent.

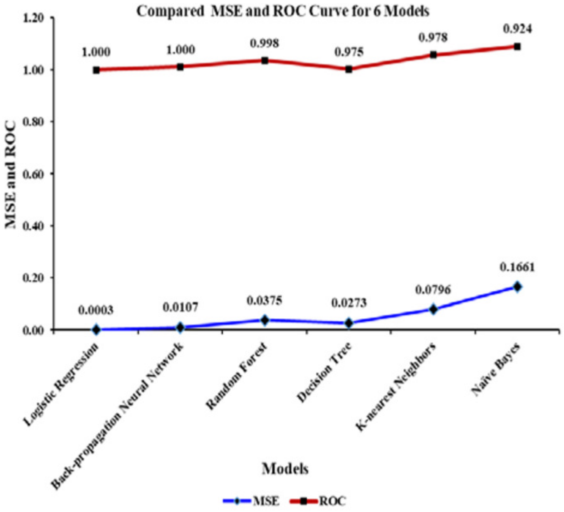

Fig.6: *Compared MSE and ROC.*

Table 5: *Confusion Matrix for Logistic Regression.*

		Predicted value				
		R1	R2	R3	R4	R5
Actual value	R1	15038	2	0	0	0
	R2	2	8256	2	0	0
	R3	0	1	4054	1	0
	R4	0	0	4	2059	4
	R5	0	0	0	3	2503

Table 6: *Confusion Matrix for Back-propagation Neural Network .*

		Predicted value				
		R1	R2	R3	R4	R5
Actual value	R1	14964	76	0	0	0
	R2	143	8061	56	0	0
	R3	0	77	3924	55	0
	R4	0	0	82	1919	66
	R5	0	0	0	50	2456

Table 7: *Confusion Matrix for Random Forest.*

		Predicted value				
		R1	R2	R3	R4	R5
Actual value	R1	14839	201	0	0	0
	R2	217	7846	197	0	0
	R3	0	211	3722	123	0
	R4	0	0	184	1773	110
	R5	0	0	6	147	2353

6. CONCLUSIONS

The objective of this study was to create a suitable model for cardiovascular disease risk screening using data from patients from 138 hospitals in Saraburi during the years 2018 to 2019. The results reveal that the logistic regression model is the best choice and has suitable performance over all. It achieves an accuracy rate of 99.940% for this data set. The accuracy value is very high. This might be because of our data preprocessing stage where all incomplete records were deleted from the study.

Moreover, most input variables are numeric which is suitable for the two top performing algorithms. To compare with Nai-arun and Moungmai [16], the random forest algorithm is not the best performance for this data set, where most input variables are numeric. However, the accuracy of the model is still very high, more than 95%. This might be because the model is suitable for dichotomous output variables and qualitative input variables.

In further work, we plan to modify the prototype model so that ensemble approaches will be applied. In addition, a variety of datasets from other areas and other factors that are related to the disease risk might be studied for developing new models.

ACKNOWLEDGEMENTS

First of all, we would like to thank to Saraburi Provincial Health Office and National Health Security Office, District 4, Saraburi province, for allowing us to use this data set. We would also like to thank all staff of these offices for consulting with us and making all information clear.

References

[1] World Health Organization, Noncommunicable Diseases, [Online]. Available: `https://www.who.int/news-room/fact-sheets/detail/noncommunicablediseases`. [17-January-2019].

[2] World Health Organization, Cardiovascular Disease, [Online]. Available: `https://www.who.int/Cardiovasculardiseases/en/`. [25-January-2019].

[3] Ministry of Public Health, Cardiovascular Disease, [Online]. Available: `https://www.moph.go.th/` [19-February-2019].

[4] Bureau of Non Communicable Disease, The issue of the World Heart Day campaign, 2018, [Online]. Available: `http://thaincd.com/document/file/download/knowledge/` [10-April-2019].

[5] P-N. Tan, M. Steinbach and V. Kumar, *Introduction to Data Mining*, Addison Wesley, USA, 2006.

[6] I. H. Witten and E. Frank, Data Mining: *Practical Machine Learning Tools and Techniques*, 2nd ed, Morgan Kaufman, USA, 2005.

[7] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed, Morgan Kaufman, USA, 2012.

[8] P. Sittidech, and N. Nai-arun, "Random Forest Analysis on Diabetes Complication Data," in *International Conference Biomedical Engineering (BioMed)*, pp.315-320, 2014.

[9] S. Dua, "Data mining and fusion paradigms of clinical informatics," in *Health data standards: From reimbursement to clinical excellence*, Bangkok, Mahidol University, pp.107-117, 2011.

[10] S. Chakrabarti, E. Cox, E. Frank, R. Güting, J. Han, X. Jiang, M. Kamber, S. Lightstone, T. Nadeau, R. E. Neapolitan, D. Pyle, M. Refaat, M. Schneider, T. Teorey, I. Witten, *Data Mining: Know It All*, Morgan Kaufmann, USA, 2008.

[11] D-Y. Yeh, C-H. Cheng and Y-W. Chen, "A predictive model for cerebrovascular disease using data mining," *Journal of Expert System with Application*, Vol. 38, pp.8970-8977, 2011.

[12] U. Suksawatchon, J. Suksawatchon, and W. Lawang, "Health Risk Analysis Expert System for Family Caregiver of Person with Disabilities using Data Mining Techiques," *International Journal of ECTI Transactions on computer and Information Technology*, Vol. 12, No. 1, pp.62-72, 2018.

[13] N. Rachata, W. Rueangsirarak, C. Kamyod, and P. Temdee, "Fuzzy-based Risk Prediction Model for Cardiovascular Complication of Patient wih Type 2 Diabetes Mellitus and Hypertension," *International Journal of ECTI Transactions on computer and Information Technology*, Vol. 13, No. 1, pp.41-50, 2019.

[14] P, K. Saxena. R. Sharma, "Efficient Heart Disease Prediction System," *International Journal of Procedia Computer Science*, 85, pp.962-969, 2016.

[15] R. Assari, P. Azimi and M. R. Taghva, "Heart Disease Diagnosis Using Data Mining Techniques," *International Journal of Economics & Management Sciences*, Vol. 6, pp.101-105, 2017.

[16] N. Nai-arun, and R. Moungmai, "Comparison of Classifiers for the Risk of Diabetes Prediction," *International Journal of Procedia Computer Science*, 69, pp.132-142, 2015.

[17] M. H. Dunham, *Data Mining, Introductory and Advanced Topics*, Prentice Hall, USA, 2002.

[18] M. T. Jones, "Artificial Intelligence: A Systems Approach," *Infinity Science*, Hingham, 2008.

[19] Towards Data Science, Machine learning fundamentals (II): Neural networks, [Online]. Available: `https://towardsdatascience.com/machine-learning-fundamentals-ii-neural-networksf1e7b2cb3eef?`. [13-March-2019]

[20] K. Kitbumrungrat, "Multinomial Logistic Regression Model for Learning Classification and Ordinal Logistic Regression Model for Student Grade Analysis," *Varidian E-Journal of Science and Technology Silpakorn University*, Vol. 4, No. 2, pp.19-35, 2017.

[21] Saraburi Provincial Health Office, Cardiovascular Disease, [Online]. Available: `http://www.sro.moph.go.th/ewtadmin/ewt/saraburi\_web/main.php?`. [11-May-2019].

[22] Ministry of Public Health, *Guidelines for Assessment of Cardiovascular Risk*, The War Veterans Organization., Bangkok, 2019.

Nongyao Nai-arun received a B.Ed. degree in Computer Education from Nakhon Ratchasima Teacher College, Nakhon Ratchasima, Thailand in 1992, and M.Sc. degree in Technology of Information System Management from Mahidol University, Bangkok, Thailand in 2000, she is a Ph.D. in Computer Science from Naresuan University, Phitsanulok in 2014. She is currently an assistant professor in the Department of Information Technology, Faculty of Science and Technology, Phranakhon Si Ayutthaya Rajabhat University. Her research interests include data mining, forecasting, artificial intelligence and machine learning.

Rungruttikarn Moungmai received her B.Sc. and M.Sc. degrees in Applied Statistics from King Mongkut's Institute of Technology North Bangkok, Thailand in 1999 and 2002 respectively and her Ph.D. in Applied Statistics from University of Reading, England in 2013. Her current research interests include Operations Research, Regression Analysis, Forecasting and Time Series Analysis, Deep Learning, Machine Learning and Imbalanced Data.