International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V69I6P217
Volume 69 Issue 6, 115-120, June 2021

Stroke Patients Classification Using Resampling Techniques and Decision Tree Learning

Sumitra Nuanmeesri[#1], Wongkot Sriurai[*2], Nattanon Lamsamut[*3]

[#] *Faculty of Science and Technology, Suan Sunandha Rajabhat University, Bangkok, Thailand*

[*] *Faculty of Science, Ubon Ratchathani University, Ubon Ratchathani, Thailand*

[1]sumitra.nu@ssru.ac.th, [2] wongkot.s@ubu.ac.th, [3] nattanon.ssktc@gmail.com

Abstract *- Stroke is a major global and worldwide public health problem. It is a major cause of mortality, morbidity, and disability in developed and increasingly in less developed countries. The goal of this study is to develop a classification model for stroke patients towards the application of resampling techniques together with the decision tree learning methods. Since the size of the collected dataset to construct the model was small, the research team applied the resampling techniques to solve the problem. When the datasets of predicted outputs were imbalanced, the data size needs balance adjustment between 100%-300%. Afterward, decision tree learning was applied to the construction of the classification model for stroke patients by which the results from three decision tree learning methods, including ID3, C4.5, and Random Forest, were compared. The model's effectiveness was evaluated by 10-fold cross-validation. The evaluation results showed that the model tested with 10-fold cross-validation and adjusted by resampling to 200% with the Random Forest technique provided the highest level of effectiveness, with the classification accuracy of 96.40%, precision of 96.45%, and recall of 96.60%. This model gave higher efficiency than the results gained from both ID3 and C4.5 techniques.*

Keywords — *Cerebrovascular Disease, Decision Tree, Patient Classification, Resampling, Stoke*

I. INTRODUCTION

Stroke is a worldwide public health issue, which the World Stroke Organization (WSO) reported as the second most common cause of death globally. Currently, there are 17 million stroke patients around the world, and 6.5 million people were killed by stroke each year on average. In Thailand, stroke is the first leading cause of premature death among women and the second for men. According to the Strategy and Planning Division, Ministry of Public Health, the average annual deaths caused by stroke per 100,000 population between 2014-2016 were 38.63, 43.28, and 43.54, respectively. This indicates that the death rates caused by stroke kept rising every year, and it also caused more deaths than diabetes or arteriosclerosis by 1.5 times or twice [1]. Many previous studies revealed that stroke patients were often admitted to the hospital unconsciously or with neurological symptoms such as hemiplegia. They were rarely treated by anticoagulants in time due to the delays of services and unavailability of medical specialists at provincial hospitals. This resulted into ineffective treatments and a higher death rate (9-10% of the total patient population). Family members, who had to hire nursing assistants throughout the patients' entire lifetime, also inevitably faced financial burdens.

Therefore, it is extremely crucial to observe the symptoms of the potential stroke patients before transporting them to the hospital in order to reduce the death rates among the transferred patients. For this reason, the research team decided to construct a model for classifying stroke patients towards the application of resampling techniques and decision tree learning. At present, data mining has been applied to classification and prediction in many fields, including education and business decision-making, and patient classification. For example, it has been used for categorizing hepatitis patients [2] and cardiovascular disease patients [3]. Decision tree learning is another data mining method regarding classification and prediction commonly used by many researchers. For example, Hongboonmee and Sornroong [4] applied the decision tree learning method to classify diseases in water buffalo on mobile phones; their research findings showed that the Random Forest algorithm provided the most effective results with an accuracy of 99.47%. Another study conducted by Mohapatra and Mohanty also explored the application of decision tree learning to data classification in the UCI database [5].

According to the above-related research studies, it is found that most studies favor the application of decision tree techniques to classify medical information. The results of the classification using this technique yield good results. However, most of the data used for classification faced information problems. These data collected will contain a small number of examples of each class, resulting in an imbalanced dataset, which reduces the classification efficiency. Therefore, to against this problem, the researchers applied the Resampling technique to address the imbalanced data problem. The well-balanced data is then used for further classification by the decision tree technique. This method will help balance the dataset and improve the efficiency of data classification more accurately.

The rest of the paper is organized as follows: section II explains the literature review, section III explains the research methodology. The research findings are presented in section IV.

II. LITERATURE REVIEWS

A. Stoke

Stroke or cerebrovascular accident (CVA) is caused by the lack of blood flow to the brain owing to blockage or rupture of an artery to the brain. Consequently, brain cells up to 2,000,000 neurons are destroyed [6], leading to a variety of symptoms. Stroke can be categorized into the following types [7]:

***a) Ischemic Stroke*:** is a kind of stroke making up 80% of its patient population. It is caused by blockage of an artery which results into a lack of blood flow. It is often accompanied by atherosclerosis, which arises from the accumulation of fats in and on artery walls. This kind of stroke can be further divided into two sub-categories as follows.

- Thrombotic Stroke is a kind of ischemic stroke caused by atherosclerosis arising from hyperlipidemia.
- Embolic Stroke is a kind of ischemic stroke caused by blockage of an artery.

***b) Hemorrhagic Stroke*:** is a type of stroke resulting from intracranial hemorrhage, which leads to bleeding within the skull. This kind of stroke is rarer than the first type, as it takes up around 20% of the stroke patient population. This kind of stroke can be further divided into two sub-categories as described below.

- Aneurysm is a type of hemorrhagic stroke caused by weakness in the arterial wall.
- Arteriovenous Malformation is a type of hemorrhagic stroke caused by innate malformation of blood vessels in the brain.

The common symptoms of stroke are listed below [7]:

- Muscle weakness or paralysis of a body part, which often happens to one side of the body, such as facial palsy and hemiplegia.
- Numbness or loss of sensation of a body part, which often happens to one side of the body.
- Difficulty speaking, for example, inability to speak, speech impediment, or inability to understand verbal communication.
- Vestibular balance problems such as losing balance while walking or sudden dizziness.
- Partial vision loss or double vision.

B. Data Mining

Data mining is a process of identifying patterns and relationships in large datasets. Nowadays, data mining is widely used in many areas, including business management, science and medicine, economics, and social research. Data mining is an evolution of data storage and interpretation. It has transformed basic data storage into a database where information can be retrieved as well as the beneficial knowledge embedded in each dataset [8], [9]. This research applies the data mining technique, including decision tree and Random Forest.

C. Decision Tree Learning

Decision tree learning is a supervised learning [10] method providing outputs that resemble a tree's structure. The target variables are grouped by the comparison of their attributes in the form of a tree-like structure where final nodes in a decision path share the same label. Each decision tree consists of root nodes (in which each node is a "test" on an attribute), branches (the possible outcomes of the test), and leaf nodes at the bottom (representing classification or decision) [11]. A decision tree can be constructed in various ways. In this study, three decision tree learning techniques are compared. These techniques consist of the Iterative Dichotomiser 3 algorithm, C4.5 algorithm, and Random Forest.

D. Iterative Dichotomiser 3 Algorithm

In decision tree learning, Iterative Dichotomiser 3 (ID3) algorithm is employed for generating a decision tree from a dataset based on the Information Theory; the estimated values are used for considering the variables for data classification. A decision tree is constructed by the selection of variables in order of the gain values (from the highest to the lowest gain values, respectively) [9], [11]. To illustrate, when considering two datasets, P and N, the number of samples in P class = p while the number of samples in N class = n; the values of datasets are the entropy or the expected number of bits of information contained in an attribute used for classifying P and N classes, as shown in (1).

$$I(p,n) = -\frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right) \quad (1)$$

Entropy is the values gained from the classification of *A* attribute, in which *A* is a regular attribute which divides s into $\{s_1, s_2, \ldots, s_v\}$ while s_1 contains P_1 from *P* class and n_1 from *N* class, as shown in (2).

$$E(A) = \sum_{i=1}^{v} \frac{p_1+n_1}{p+n}(p_1,n_1) \quad (2)$$

Thus, the Data Gain obtained from the classification based on *A* regular attribute is provided, as demonstrated in (3).

$$Gain(A) = I(p,n) - E(A) \quad (3)$$

E. C4.5 Algorithm

C4.5 algorithm is an extension of ID3 algorithm, which is used for data classification. It generates a decision based on gain ratios instead of gains by selecting the most significant attributes (those with the highest gain ratios) as the root nodes. Then, the Entropy, Information Gain, and Split Information are calculated [9], [11]. Entropy is calculated by measuring information as demonstrated in (4).

$$Entropy(s) = \sum_{i=1}^{c} -P_i \log_2 P_i \quad (4)$$

Where:

S is the attribute used for measuring entropy;

P_i is the proportion of members in *i* class, which is equal to the total number of sampling members.

Information Gain is calculated by identifying the information gain before the standard values or gain ratios, as shown in (5).

$$Gain(S,A) = Entropy(s) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (5)$$

In which,

A refers to A attribute;

$|S_v|$ refers to A attribute with V value;

$|S|$ refers to the number of sampling members.

Split Information is calculated by splitting the data according to the value of the attributes, as demonstrated in (6).

Split Information is calculated by splitting the data according to the value of the attributes, as demonstrated in (6).

$$Spilt\ Information(S,A) = -\sum_{i=1}^{n} \frac{|S_i|}{S} \log_2 \frac{|S_i|}{|S|} \quad (6)$$

In which, S_i refers to the proportion of the number of i members.

Gain Ratio is an extension of ID3 algorithm used for reducing a bias towards multi-valued attributes, as shown in (7).

$$Gain\ Ratio(S,A) = \frac{Gain(S,A)}{Split\ Information(S,A)} \quad (7)$$

F. Random Forest

Random Forest (RF) is a decision tree learning method to increase the diversity of a model by randomizing attributes and then replacing the samples. Some samples are excluded or "Out-of-Bag (OOB)" and then tested. This method is called "Bagging" and used for finding the most voted outputs [12].

The principle of Random Forest is to create a model based on multiple decision tree sub-models. In each model are different datasets; they are the subsets of the whole dataset. When processing the prediction, each decision tree will independently predict its results. The prediction calculation relies on the voting of most selected outputs by the Decision Tree technique (in case of classification). The advantages of this approach are the accurate prediction results and a low rate of overfitting problems [12].

Figure 1 shows the Random Forest process starts from sampling the data (bootstrapping) from the entire dataset in order to create n independent datasets based on the number of decision trees in Random Forest. For example, suppose the original dataset may have nine features (X_1, X_2, …, X_9). Each decision tree has different features, but not every row has complete information from the dataset (X_1 -> X_1', X_2->X_2', …, X_9->X_9'). Next, the decision tree models for each dataset are created. Then, the outputs of each model are aggregated (bagging), for instance, voting in case of classification.

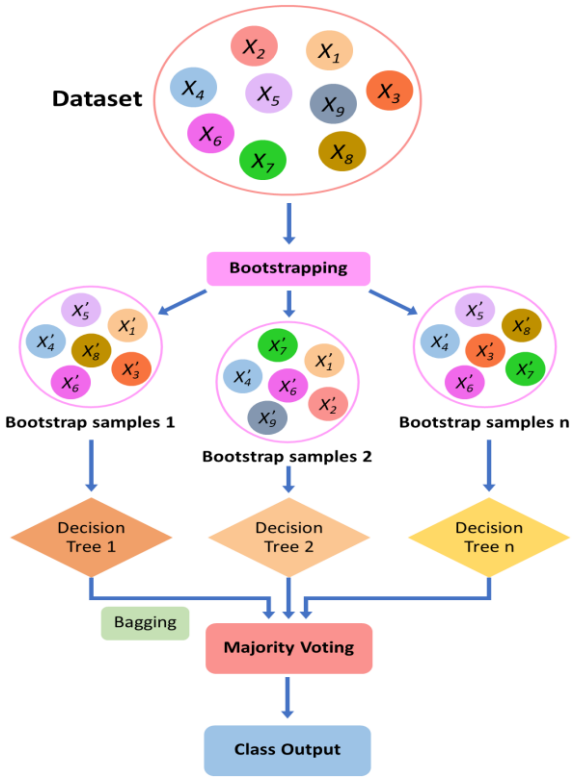

Fig. 1 Principle of Random Forest

III. METHODOLOGY

This research proposes the application of resampling techniques together with decision tree learning methods to the classification of stroke patients. The research process consists of four stages: 1) data collection, 2) data preparation for data mining, 3) model development, and 4) model's effectiveness evaluation. The overall research process could be illustrated in Fig. 2.

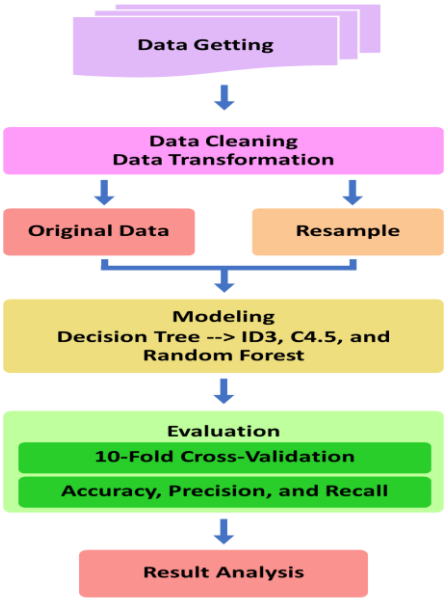

Fig. 2 Research Process

A. Data Collection

In this study, data used for exploring and classifying stroke patients was collected from the patient delivery database of a hospital. The data was collected between 2014 and 2018, and the datasets included 741 records. The data contains attributes gained from external physical check-ups, medical diagnosis, and specific medications given to the admitted stroke patients.

B. Data Preparation for Data Mining

The attributes used for constructing the model were collected from stroke patient records, which have been through the data cleaning and data transformation processes. These attributes are listed in Table I.

TABLE I
ATTRIBUTES USED FOR THE MODEL CONSTRUCTION

No.	Attributes	Description
1	loads_id	Patient delivery method codes based on "LOADS" matrix 1 = Ambulance 2 = Ambulance with Nurses 3 = Ambulance with Physicians and Nurses 4 = By Patient 5 = EMS Officers
2	is_receive	Original service providers canceled online patient data transfer. 0 = Not Received, 1 = Received
3	is_referboard	Displayed on Refer Board Y = Displayed on Refer Board N = Not Displayed on Refer Board
4	stroke_f	F (Bell's Palsy) Y = Yes, N = No
5	stroke_a	A (Arm or Leg Weakness) Y = Yes, N = No
6	stroke_s	S (Difficulty Speaking) Y = Yes, N = No
7	stroke_t	T (Treated within 4.5 Hours) Y = Treated within 4.5 Hours N = Treated beyond 4.5 Hours
8	treat_iv	Treated by On IV Fluid Y = Yes, N = No
9	treat_ekg	Treated by EKG Y = Yes, N = No
10	treat_ct	Treated by CT Y = Yes, N = No
11	treat_fc	Treated by Folay's Cath Y = Yes, N = No
12	treat_ng	Treated by NG-TUBE Y = Yes, N = No
13	treat_et	Treated by ET-TUBE Y = Yes, N = No
14	treat_tt	Treated by TT-TUBE Y = Yes, N = No
15	treat_o2	Treated by Oxygen Therapy (8 L/min via Face Mask) Y = Yes, N = No
16	cannula_o2	Treated by Oxygen Therapy Y = Yes, N = No
17	treat_other	Other kinds of Treatment Y = Yes, N = No
18	class	Yes = Have a stroke No = Not have a stroke

C. Model Development

Once the data transformation process had been completed, the research team cross-checked the reliability of the data and then analyzed it. The research team found that the number of some datasets was too small, so the data imbalance was adjusted by resampling techniques. Resampling was a data imbalance resolution method used when the predicted outputs were imbalanced. In this study, the values were set as follows: the random seed = 5; noReplacement = false; bias factor towards uniform class distribution = 0.0; sample size percentage = 100-300. After adjusting the data imbalance by resampling, the adjusted datasets were employed to develop a model with three decision tree learning methods, which included ID3, C4.5, and Random Forest.

D. Model's Effectiveness Evaluation

In this research, the model's effectiveness was evaluated by 10-fold cross-validation to allow all datasets to be both training data and testing data potentially. The datasets were equally divided into two groups of data: the first group was training data, while the second group was testing data. They have switched around ten times to evaluate the model's effectiveness by comparing the accuracy [13], [14] precision, and recall values gained from each decision tree learning method.

IV. RESULTS

The model's effectiveness evaluation results were processed by the program so-called "Weka" version 3.9. The evaluation method was 10-fold cross-validation. The data used for constructing the model had been through resampling in order to adjust its imbalance. Next, the adjusted data was classified by three decision tree learning methods, including ID3, C4.5, and Random Forest. The 10-fold cross-validation approach was used to test the model's effectiveness by comparing the accuracy, precision, and recall values gained from each decision tree learning method. The evaluation results are illustrated in Table II.

TABLE II
THE MODEL EVALUATION RESULTS

Algorithm	Dataset	Data (rows)	Accuracy (%)	Precision (%)	Recall (%)
ID3	Original	741	89.50	89.40	89.65
C4.5	Original	741	91.45	91.50	91.70
Random Forest	Original	741	94.50	94.35	94.60
Resample & ID3	100%	1,110	91.60	91.55	91.80
Resample & ID3	200%	1,480	92.15	92.20	92.40
Resample & ID3	300%	2,220	91.35	91.25	91.50
Resample & C4.5	100%	1,110	91.40	91.55	91.60
Resample & C4.5	200%	1,480	92.30	92.35	92.50
Resample & C4.5	300%	2,220	91.45	91.65	91.70
Resample & Random Forest	100%	1,110	95.25	95.40	95.50
Resample & Random Forest	200%	1,480	96.40	96.45	96.60
Resample & Random Forest	300%	2,220	95.40	95.60	95.50

According to Table II, the model tested by 10-fold cross-validation, with data imbalance adjustment by resampling techniques, provided the highest level of effectiveness when the data was classified by the Random Forest technique. The model using Random Forest for data classification provided the accuracy of 96.40% shown as Fig. 3, the precision of 96.45% shown as Fig. 4, and the recall of 96.60% shown as Fig. 5, which was higher than the results given by ID3 or C4.5 shown as Fig. 6.

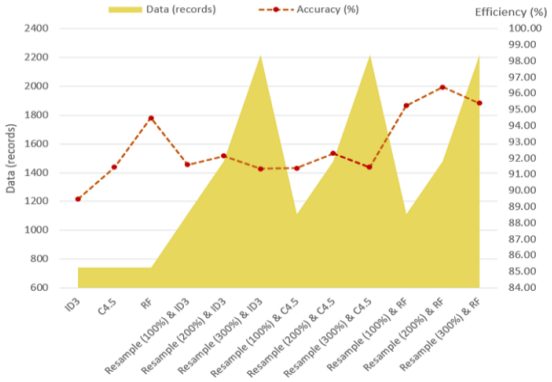

Fig. 3 Comparison of the Accuracy of the Model's Effectiveness Evaluation

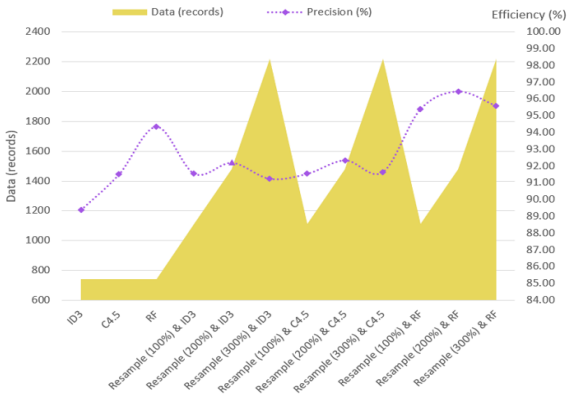

Fig. 4 Comparison of the Precision of the Model's Effectiveness Evaluation

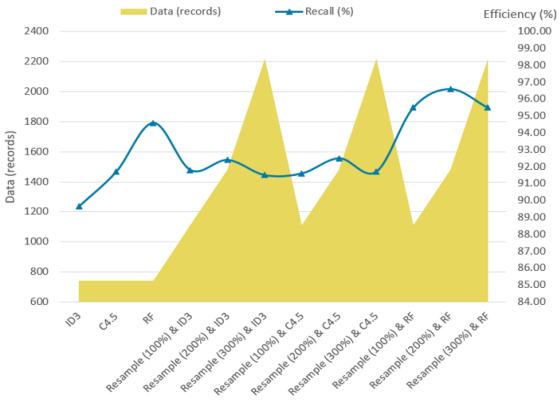

Fig. 5 Comparison of the Recall of the Model's Effectiveness Evaluation

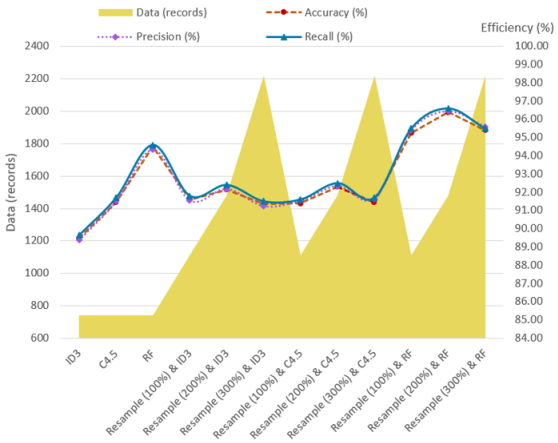

Fig. 6 Comparison of the Model's Effectiveness Evaluation

V. CONCLUSIONS

This research develops a model for stroke patients' classification towards the application of resampling techniques and decision tree learning methods. The data used for developing the model had been through the data imbalance adjustment process based on the resampling approach. The data imbalance adjustment was conducted between the range of 100%-300%, and it was found that increasing the data size to 200% provided the highest level of effectiveness. Once the data imbalance adjustment process had been completed, the dataset was used for developing the model with three decision tree learning methods, including ID3, C4.5, and Random Forest. The research findings revealed that the model tested by 10-fold cross-validation, with data imbalance adjusted by resampling, provided the most effective classification results when it employed Random Forest for data classification; the model using the Random Forest technique provided the accuracy of 96.40%, which was higher than of the model using ID3 or C4.5. Therefore, the developed model that applied the Random Forest technique was found to be the most applicable model for developing a stroke patient classification mobile application. When a loaded patient is checked up or diagnosed during the pre-hospital period and is labelled with either having or not having a stroke, an alert will be sent to the destination hospital in order to prepare equipment and staff required for treating a stroke patient. This study's findings conform to the research conducted by Mohapatra and Mohanty [5], which suggested that the application of Random Forest to data classification could increase the effectiveness of data classification.

The research findings revealed that the application of resampling to data imbalance adjustment could effectively adjust the data imbalance. After the data was balanced, the data was delivered to the data categorization process. This resulted into more effectiveness in data categorization. Future research should apply "image processing" to analyzing data because it can provide more accuracy, effectiveness, and convenience to the data analysis process.

ACKNOWLEDGMENT

The researchers are grateful to the Institute for Research and Development, Suan Sunandha Rajabhat University, and the Faculty of Science at Ubon Ratchathani University, who supported and gave this research opportunity.

REFERENCES

[1] Bureau of Non-Communicable Diseases, World stroke day campaign keystones, Department of Disease Control, (2017) 1–4.

[2] S. Uhmn, D. Kim, S. W. Cho, J. K. Cheong, and J. Kim, Chronic hepatitis classification using SNP data and data mining techniques, in Proc. Frontiers in the Convergence of Bioscience and Information Technologies, Jeju, South Korea, (2007) 81–86.

[3] G. Choudhary and S. N. Singh, Prediction of cardiovascular disease using data mining technique, in Proc. 4th International Conf. Information Systems and Computer Networks, 2019, pp. 99–103.

[4] N. Hongboonmee, P. Sornroong, Applying decision tree classification techniques for diagnose the disease in cow on mobile phone, Journal of Science and Technology, Ubon Ratchathani University, 20 (1) (2018) 44–58.

[5] S. K. Mohapatra, Mohanty, Analysis of resampling method for arrhythmia classification using random forest classifier with selected features, in Proc. 2nd International Conf. Data Science and Business Analytics, Changsha, China, (2018) 495–499.

[6] N. S. Kumar, M. Thangamani, V. Sasikumar, and S. Nallusamy, An improved machine learning approach for predicting ischemic stroke, International Journal of Engineering Trends and Technology, 69(11) (2021) 111–115.

[7] P. Pisuttakoon. Stroke [Online]. Available: http://www.med.nu.ac.th/dpMed/fileKnowledge/106_2017-08-19.pdf. (2018).

[8] J. Han, and M. Kamber, Data Mining Concepts and Techniques, U.K.: Morgan Kaufmann Publishers, (2011).

[9] P. N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, Introduction to Data Mining, 1st ed., NJ: Pearson, (2014).

[10] Aman and R. S. Chhillar, Disease predictive models for healthcare by using data mining techniques: State of the art, International Journal of Engineering Trends and Technology, 68 (10) (2020) 52–57.

[11] C. Kaewchinporn, Data classification with decision tree and clustering techniques, Thesis in Computer Science, King Mongkut's Institute of Technology Ladkrabang, Thailand, (2010).

[12] L. Breiman, Random forests, Machine Learning, 45 (2001) 5–32.

[13] S. Nuanmeesri, Development of community tourism enhancement in emerging cities using gamification and adaptive tourism recommendation, Journal of King Saud University - Computer and Information Sciences, (in press), (2021).

[14] S. Nuanmeesri and W. Sriurai, Thai water buffalo disease analysis with the application of feature selection technique and Multi-Layer Perceptron Neural Network, Engineering, Technology & Applied Science Research, 11 (2021) 6907–6911.