Interdisciplinary Research Review
Volume 14, Number 4, Pages 61 – 69

Data mining model and application for stroke prediction: A combination of demographic and medical screening data approach

Sotarat Thammaboosadee[1,*] and Teerapat Kansadub[2]

[1]**Information Technology Management Division, Faculty of Engineering, Mahidol University, Thailand**
[2]**Faculty of Physical Therapy, Mahidol University, Thailand**

Abstract

This paper presents the data mining process that was used for building a stroke prediction model based on demographic information and medical screening data. The data that was gathered from a physical therapy center in Thailand comprised of outpatients' medical records, medical screening forms, and a target variable. A group of 147 stroke patients and 294 non-stroke individuals with six demographic predictors were selected for the study. Three classification algorithms were used in the study. These were; Naïve Bayes, Decision Tree, and Artificial Neural Network (ANN). They were used to analyze the data collected and the results were compared. They were evaluated by use of a 10-fold cross-validation method. The selection criteria were primarily measured by accuracy and the area under ROC curve (AUC). The secondary selection criteria were indicated by False-Positive Rate (FPR) and False-Negative Rate (FNR). The results showed that the best performing algorithm that was studied was ANN combined with integrated data. This approach have an overall accuracy of 0.84, an AUC of 0.90, a FPR of 0.12 and an FNR of 0.25. The results of the study demonstrated that ANN with the integration of demographic and medical screening data produced the best predictive performance compared to the other models. This result was found according to both the primary and secondary model selection criteria.

Keywords: Stroke prediction, data mining, medical screening data
Article history: Received 13 February 2019, Accepted 30 August 2019

1. Introduction

Stroke is a disease that affects the arteries leading to the brain [1]. It would cause the abnormalities of vascular in a brain and affecting of the nerves such as muscle weakness, numbness and can be fatal. It can be separated into two types: Ischemic stroke and Hemorrhagic stroke. Ischemic stroke occurs as a result of an obstacle within a blood vessel transferring blood to the brain. The underlying condition for this type of obstacle is the development of fatty deposits lining the vessel walls. The common consequence includes aphasia, physical disability, losing of cognition, communication skills, depression and other mental health problems. Moreover, stroke is a major public healthcare concern and has a significant impact on individuals, families and wider society. Recently, the World Health Organization [2] reported that stroke is the third leading cause of mortality overall life periods.

Stroke identification is tedious and time-consuming for medical diagnosis which is initially driven by experts' experience. Therefore, it would be beneficial if there is an automated system to predict the risk of stroke in patients. There are various types of data involved in the analysis depending on availability and level of abstraction. Therefore, those methods are to identify the independent risk factors based on the certain data source, some of them employed several input data such as medical history, symptoms, and the theoretical proven to be accurate risk factors. However, those secondary data may be difficult to collect and need special medical equipment. The high level of abstraction is required to discover related risk factors and build the complex identification model. Furthermore, some data analytics technology are required for creating the model. One analytical technology, data mining, is known for the knowledge discovery from the database [3]. It is an interdisciplinary field to discover patterns or model in high-volume data involving methods from several areas, such as artificial intelligence, machine learning, statistics, and database systems [4].

Some prior studies related to the finding of the stroke risk factors are reviewed. Preliminary, study was made to discover stroke risk factors, specifically

*Corresponding author; email:sotarat.tha@mahidol.ac.th

for Thai citizen [5]. This study found that the factors are Hypertension, Diabetes, Obesity, Dyslipidemia, Smoking, Cardiovascular disease (CAD) and Drug use, in descending order of severity. Hanchaiphiboolkul *et al*. [6] proposed another stroke risk factors study. They applied cross-sectional analysis by using baseline health survey data which are age, gender and region to discover stroke prevalence in Thailand. By using multiple logistic regression as a tool based on 19,997 subjects, the study found that stroke prevalence in Thailand has few continuously increasing but lower than other developed countries. In geographic variation, stroke prevalence mostly discovered in Bangkok, Central, and Southern regions. Moreover, the age over 65 years old, male gender, and occupation class are found to be slightly significant. This study motivates us in model building aspect by using the demographic data as the primary source.

The most recent work proposes Arslan *et al*. [7 aimed to develop the medical data mining processes for extracting patterns approaching to predict ischemic stroke from collected dataset belonging to Inonu University, Turkey. The experiments were compared among three classifiers: Support Vector Machine (SVM), Stochastic Gradient Boosting (SGB), and Penalized Logistic Regression (PLG) which worked on 80 stroke patients and 112 healthy individuals consisting of demographic data and blood test results. The experiment showed that the SVM was the best classifier indicated by 0.9789 accuracies and 0.9783 AUC. This finding supports our hypothesis that the stroke risk prediction model is possible to create. Moreover, the important variables were also explored that the top three essential factors were age, Creatinine (CR), and Chlorine (CL), consecutively. Interestingly, the gender and marital status were slightly relevant in the eighth and thirteenth from overall seventeen variables. This work proved that the demographic data could be useful in model building.

Amini *et al*. [8] studied the stroke prediction based on data mining model with the data collected from Iran's hospital. The 50 risk factors of 807 patients dataset was experimented by two prediction algorithms: k-Nearest Neighbors and C4.5 Decision Tree. The results showed that the C4.5 decision tree performed best accuracy and precision. However, those risk factors were excessive and impossible to gather by end users who do not own a medical equipment and may cause difficulties in data collection. Sudha *et al*. [9] also proposed an alternative stroke prediction model based on medical history and symptoms including physical exam results, blood test results and diagnoses. The test data were collected from medical institute. Therefore, the data preprocessing was performed by removing duplicate records, missing data, noisy and inconsistency. They compared the model by four indicators which are accuracies, false-positive, false-negative, and AUC, combined with three classification methods: C4.5 Decision Tree, Naïve Bayes, and Artificial Neural Network (ANN). The result showed that the C4.5 decision tree is the best classification algorithm proven by 0.98 accuracies. In the real application, plenty of inputs may confuse the non-medic users. Thus, use of more concise input data is challenging for developing the prediction model. Nevertheless, their work has motivated us towards selecting understandable data as an input set and setting up of the experiment is interesting and may be applied in our work. Other prior related works on stroke prediction model were also reviewed.

Other related studies on the application of data mining in stroke disease are widely developed. Easton *et al*. [10] examined the risk factors of short-term and short/intermediate-term for post-stroke mortality using Naïve Bayes, Logistic Regression, and Decision Tree. Panzarasa *et al*. [11] used the classification tree for analyzing the stroke care process which aimed to identify the specific key indicators and was able to monitor the quality of medical process for stroke care.

From abovementioned research, the data mining have been widely proposed in various aspects for the stroke including data source, data schema, types of prediction models and applications. However, most of them are based on the factors that are inconvenient for end users to obtain. In this paper, we develop the stroke prediction model based on two kinds of outpatients' information, which are demographic data and medical screening data. Both data types are collected from a physical therapy center in Thailand during 2012 to 2015. The demographic data has been electronically stored in the relational database while the medical screening data has been in a paper-based format, called medical screening form. The medical screening form has been designed for preliminary self-screening outpatients in the center before transferring them to the proper clinic. By using the medical screening form, the information gathered from patients is more concise and easily understood. Then, the smashing prediction model is selected from three classification data mining algorithms, namely Naïve Bayes, C4.5 Decision Tree and Artificial Neural Network (ANN). They are experimented with both information sources and their integration. Finally, an application based on the best model will be extensively demonstrated for ease of use. The expected benefit of this research is the stroke risk prediction model based on high-level of abstraction data which are demographic and medical screening data evaluated in both technical and medical aspects.

The remainder of the paper is organized as follows: the coming section introduces the research methodology including some technical issues and setting up of the experiments. The next section presents the experimental results and discussion on technical issues, some insights into the medical domain and demonstration of the system application. Finally, we conclude

Parameter setting of modeling algorithms:

Naïve Bayes: *None*

C4.5 Decision Tree:

Minimum number of instances per leaf: 2

ANN Parameters:

Activation function: Unipolar Sigmoid Function

Number of hidden layers: 1

Initial number of hidden nodes: (attributes + classes)/2

Minimum Mean Squared Error: 0.05

Learning rate range: 0.05 to 0.5

Momentum range: 0.1 to 0.4

Evaluation method: *10-fold cross-validation*

Model Selection Criteria:

Primary: accuracy and AUC
Secondary: FPR and FNR

Figure 1: The setting up of the experiments.

the paper and suggest some directions for future research.

2. Methodology

2.1. Data source

The required information for this study is data of all outpatients during 2012 to 2015. In our study, there are two data types in this study: demographic data and medical screening data which will then be separately described for their characteristics, collection procedures and preprocessing method in the next subsections.

2.2. Demographic data collection and preprocessing

The demographic data refers to the general information of clients such as H/N number, name, gender, occupation, etc. This kind of information is already stored in the accessible relational database and can be gathered by database querying. An original demographic data source consists of 8 tables with total 112 attributes and more than 100,000 transactions. This high volume of data is needed for selection and filtering procedure. A general exclusion criterion is an age [5]. It was reported that age under 20 years old should be excluded in stroke determination for meaningful analysis. Moreover, since the medical diagnosis has been encoded in ICD-11 code [12] standard format, the inclusion criteria will filter the range of I60 (non-traumatic subarachnoid hemorrhage) to I69 (sequelae of cerebrovascular disease) as stroke patients and the rest is non-stroke falling in other attributes. Some demographic characteristics were eliminated due to their rare occurence. Thus, the remaining gathered factors

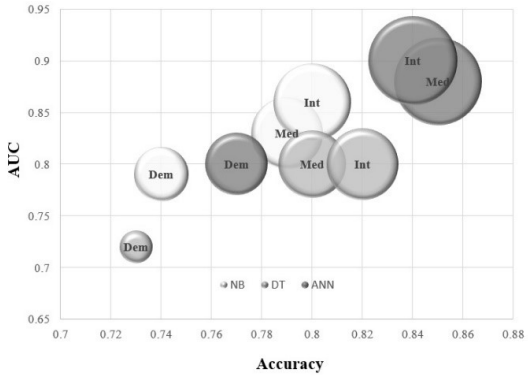

Figure 2: Comparison of accuracy and AUC of all datasets and algorithms.

are sex, age, province, marital status, education, and occupation.

Once the demographic data had been prepared already, it was found that the proportion of healthy and stroke patients revealed a vast imbalance (67,010:147) and should be resampled. Theoretically, the data resampling method can be chosen between under-sampling and over-sampling. The under-sampling is to remove majority class randomly. On the other hands, over-sampling helps to achieve a more balanced class distribution by replication minority class sample or combining it together [13]. In this research, the non-stroke patients should be down-sampled to reduce the proportion instead of using the up-sampling method because the quantity of majority class is remarkably higher than another group. After the down-sampling procedure, the final ratio between those two groups is 294:147. The characteristic, quantity and distribution of the final demographic data are shown in Table 1.

2.3. Medical Screening data collection and preprocessing

In the selected physical therapy center, a medical screening form is a self-input document which is designed to pre-filter unregistered outpatients and transfer them to the most suitable clinic possible. The medical screening form which consists of 29 simple questions with three answer choices (Yes/No/Unknown) in a paper-based format. The resampled patient records with HN number from the demographic data collection step were retrieved manually from archival storage by an assistant using spreadsheet application. During the collection process, the researcher found that the screening form was continuously improved and updated which caused four screening form versions between 2012 and 2015. Thus, some items of those forms needed to be merged into a single format as summarized in Table 2.

When the collection process of both data types was completed, we obtained three dataset for the experiment: demographic data, medical screening data and

Table 1. Preprocessed demographic data characteristics

Attributes	Values	Quantity (N=441)	Stroke	Non-Stroke
Stroke	Stroke	147		
	Non-Stroke	294		
Sex	Male	174	83	91
	Female	266	64	202
Age in Years	(Numerical)	-	67.3 (Mean)	53.90 (Mean)
Province	Capital City	302	95	207
	Fringe	109	44	65
	Countryside	25	6	19
Marital Status	Single	161	36	125
	Cohabit	246	97	149
Education	Primary	57	39	18
	High School	44	19	25
	Vocational	30	14	16
	Diploma	18	4	14
	Bachelor	156	35	121
	Master	51	8	43
	Doctoral	3	1	2
Occupation	Public servants	99	26	73
	Merchant	42	13	29
	Farmers	44	11	33
	Steward	79	33	46

their integration. The integrated dataset is the merging of both data sources using HN number as a joining key. Therefore, the experiment was conducted by datasets of 441 patients combined with three modeling algorithms described in the next step.

2.4. Modeling

In this paper, the predictive modeling is experimental compared between three classification algorithms: Naïve Bayes, Decision Tree, and Artificial Neural Network (ANN). Theoretically, each method has a unique advantage in simplification, interpretability and powerful computation. Their details including principle, strength and limitation are as following:

2.4.1. Naïve Bayes

The Naïve Bayes [14] depends on Bayes' theorem which works on the probabilistic statistical classifier. The major advantage of this method includes rapidity of use and simple for handling the dataset containing several attributes. Firstly, the dataset would be transformed into a frequency table consisting of each attributes value of all attributes. Then, the likelihood of each value is calculated by using probabilities respect to each class. When it is applied for a new case, the Naïve Bayes equation [15] is used to determine the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction. The main advantage of this method is its fast training due to the single round of database scanning. Nevertheless, its limitation is that all attributes are considered to be independent.

2.4.2. Decision tree

Another classification algorithm used in this research, which is interpretable and provides a step-by-step determination, is the decision tree. One of the well-known decision tree algorithms is C4.5 [16] and its sibling. A robust interpretable computational method of this extension of the ID3 algorithm used to generate a decision tree whose construction is based on the concept of information entropy (Quinlan 1993). Firstly, the algorithm finds an effective split of the data based on the highest normalized information gain [17] for each attribute. It then creates a decision node using the selected node and the expected value of splitting. The algorithm recurs on the split data on the selected attribute and adds these nodes as child nodes. A distinct advantage of this method is its interpretable result. However, a small variation in data can lead to different decision trees, especially in small training size.

2.4.3. Artificial neural network (ANN)

Apart from a simple tabular model like the Naïve Bayes or interpretable method like the decision tree, a meta-heuristic approach such as an ANN could be an effective classifier, particularly when the interested domain is high volume and complicated. Theoretically, the ANN is a computational model that is inspired by the structure and function of the biological neural system. It consists of an interconnection of artificial neu-

Table 2. Preprocessed demographic data characteristics

Attributes	Stroke	Quantity (N=441)		
		Attribute Values		
		Yes	No	Unknown
Hypertension	Stroke	111 (76%)	27 (18%)	9 (6%)
	Non-stroke	65 (22%)	179 (61%)	50 (17%)
Diabetes	Stroke	53 (36%)	76 (52%)	18 (12%)
	Non-stroke	18 (6%)	223 (76%)	53 (18%)
Heart disease	Stroke	29 (20%)	95 (65%)	23 (16%)
	Non-stroke	28 (10%)	248 (84%)	18 (6%)
Asthma Bronchitis Allergy	Stroke	15 (10%)	111 (76%)	21 (14%)
	Non-stroke	54 (18%)	225 (77%)	15 (5%)
Hyperlipidemia	Stroke	57 (39%)	20 (14%)	71 (48%)
	Non-stroke	34 (12%)	68 (23%)	191 (65%)
Accident	Stroke	25 (17%)	86 (59%)	36 (24%)
	Non-stroke	56 (19%)	180 (61%)	58 (20%)
Fracture	Stroke	22 (15%)	90 (61%)	35 (24%)
	Non-stroke	29 (10%)	202 (69%)	63 (21%)
Cancer	Stroke	6 (4%)	107 (73%)	34 (23%)
	Non-stroke	5 (2%)	229 (78%)	60 (20%)
Rheumatoid Gout	Stroke	21 (14%)	93 (63%)	33 (22%)
	Non-stroke	50(17%)	185(63%)	59(20%)
Tuberculosis	Stroke	2(1%)	109(74%)	36(24%)
	Non-stroke	3(1%)	228(78%)	63(21%)
Osteoporosis	Stroke	10(7%)	94(64%)	43(29%)
	Non-stroke	14(5%)	206(70%)	74(25%)
Weight Change	Stroke	7(5%)	99(67%)	41(28%)
	Non-stroke	13(4%)	215(73%)	66(22%)
Urinary Incontinence	Stroke	28(19%)	89(61%)	30(20%)
	Non-stroke	32(11%)	59(20%)	203(69%)
Vertigo	Stroke	20(14%)	49(33%)	78(53%)
	Non-stroke	19(6%)	87(30%)	188(64%)
HIV	Stroke	0(0%)	74(50%)	73(50%)
	Non-stroke	0(0%)	102(35%)	192(65%)
Liver disease	Stroke	2(1%)	68(46%)	77(52%)
	Non-stroke	6(2%)	94(32%)	194(66%)
Herpes zoster or Psoriasis	Stroke	1(1%)	71(48%)	75(51%)
	Non-stroke	5(2%)	98(33%)	191(65%)
SLE	Stroke	0(0%)	72(49%)	75(51%)
	Non-stroke	0(0%)	101(34%)	193(66%)
Depressive	Stroke	13(9%)	56(38%)	78(53%)
	Non-stroke	9(3%)	94(32%)	191(65%)
Pregnant	Stroke	0(0%)	69(47%)	78(53%)
	Non-stroke	0(0%)	102(35%)	192(65%)
Kidney	Stroke	7(5%)	105(71%)	35(24%)
	Non-stroke	3(1%)	227(77%)	64(22%)
Family Cancer	Stroke	21(14%)	81(55%)	45(31%)
	Non-stroke	62(21%)	163(55%)	69(23%)
Family Heart Disease	Stroke	27(18%)	70(48%)	50(34%)
	Non-stroke	65(22%)	154(52%)	75(26%)
Family Diabetes	Stroke	57(39%)	49(33%)	41(28%)
	Non-stroke	110(37%)	114(39%)	70(24%)
Family	Stroke Stroke	67(46%)	41(28%)	39(27%)
	Non-stroke	23(8%)	188(64%)	83(28%)
Family Heredity	Stroke	3(2%)	75(51%)	69(47%)
	Non-stroke	13(4%)	175(60%)	106(36%)

Table 2. Preprocessed demographic data characteristics (Cont.)

Attributes	Stroke	Quantity (N=441)		
		Attribute Values		
		Yes	No	Unknown
Bleed	Stroke	4(3%)	36(24%)	107(73%)
	Non-stroke	2(1%)	129(44%)	163(55%)
Muscle	Stroke	22(16%)	99(73%)	15(11%)
	Non-stroke	199(65%)	89(29%)	17(6%)
Loss Balance	Stroke	108(55%)	89(45%)	1(1%)
	Non-stroke	111(46%)	117(48%)	15(6%)

rons, and it processes information using a connectionist approach. There are several types of ANN. In this research, an ANN with a feedforward and backpropagation architecture [18] is chosen. The backpropagation algorithm learns the weights for a multi-layer network with a specified number of nodes and their connections. It aims to minimize the mean-squared error that quantifies the difference between network output values and the target values for these outputs. The ANN iteratively computes the error of the production and the gradient with respect to the error in order to update all weights in the network. The training will be eliminated when the threshold criteria such as a mean square error are reached.

There are several parameters for the training process of ANN: (1) A number of hidden layers and hidden nodes: The optimal number of units in the hidden layer of any network is difficult to determine. Specifically, in back propagation learning, there have been many reports recommending numbers of hidden layers and hidden nodes due to the generalization, complexity, and overfitting [19]. However, the study by Eberhart *et al*. [20] reported that a single hidden layer is sufficient to transform any non-linear functional relationship. Therefore, only a single hidden layer is used in this research.

(2) Learning Rate: A learning rate (theoretically a fraction between 0.0 and 1.0) will affect how quickly and efficiently the network is trained. A lower learning rate means a slower learning tempo which causes a longer training time. If this parameter is specified too high, the network may move to local optimum quickly, but it may also jump over the global optimal point which causes divergence or an oscillational effect.

(3) Momentum: This parameter (theoretically a fraction between 0.0 and 1.0) can accelerate or decelerate the learning process. The learning pace will be increased when all weight changes are going to the same direction to speed up the convergence, otherwise slower to find ways to escape from the stagnation. Too high momentum causes the learning process to slow down.

(4) Stopping criteria: This parameter specifies threshold parameter to stop the training process which is a level of residual error. The mean square error

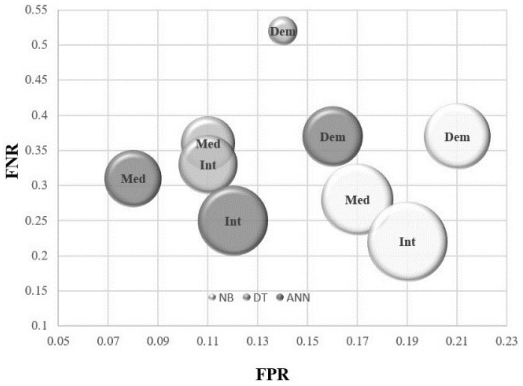

Figure 3: Comparison of FPR and FNR of all datasets and algorithms.

(MSE) is measured in every iteration of training to quantify the difference between the target and the calculated output. The training process is halted once the MSE is less than or equal to the set value. This stopping criterion is to stop the training in case the solution is divergent or oscillated.

2.5. Evaluation

In this research, the K-fold cross-validation [21] is selected to evaluate the model. In K-fold cross-validation, the original samples are randomly partitioned into K subsamples. A single subsample is held for validating. Then the cross-validation procedure repeats K times (the number of folds), with each of the K subsamples used exactly once as the validation data. Then the K results from the folds are averaged as a single outcome. The benefit of this method is that all data are used for both training and validation, and each observation is used for validation exactly once. In general, 10-folds cross validation is commonly used.

2.6. Performance measurements

In the current study, accuracy, area under the Receiver Operating Characteristic curve (AUC), false-positive rate (FPR), and false-negative rate (FNR) were utilized as model evaluation metrics [3]. These measurements are defined below in equation (1) to (3).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Table 3. Comparative experimental results.

Method	Dataset	Accuracy	AUC	FPR	FNR
NB	Dem	0.74	0.79	0.21	0.37
	Med	0.79	0.83	0.17	0.28
	Int	0.80	0.86	0.19	0.22
DT	Dem	0.73	0.72	0.14	0.52
	Med	0.80	0.80	0.11	0.36
	Int	0.82	0.80	0.11	0.33
ANN	Dem ($i = 6, h = 2, o = 2, l = 0.3, m = 0.1$)	0.77	0.80	0.16	0.37
	Med ($i = 29, h = 2, o = 2, l = 0.1, m = 0.2$)	0.85	0.88	0.08	0.31
	Int ($i = 35, h = 37, o = 2, l = 0.3, m = 0.2$)	0.84	0.90	0.12	0.25

$$FPR = \frac{FP}{FP + TN} \tag{2}$$

$$FNR = \frac{FN}{TP + FN} \tag{3}$$

where TP is the number of true-positives (correctly classified stroke patients), TN is the number of true-negatives (correctly classified non-stroke patients), FP is the number of false-positives (misclassified non-stroke patients), and FN is the number of false-negatives (misclassified stroke patients). Additionally, the area under the ROC curve (AUC) is a measurement of how well parameters can distinguish between two diagnostic groups. The AUC quantifies the accuracy regarding overall ability of the test to discriminate those patients.

2.7. Setup of the experiments

The setting up of the experimental parameters according to the information of dataset and algorithms discussed in previous subsections are shown in Figure 1.

2.8. Model deployment

A sample application is developed and deployed based on the selected model to approach the practicality. The result is demonstrated in the next section.

3. Result

3.1. Model performance

Experimental results are shown in Table 3. As introduced, the results are compared between three classification algorithms (the Naïve Bayes (NB), the C4.5 decision tree (DT), and Artificial Neural Network (ANN)) and three datasets (demographic data (Dem), medical screening data (Med), and integrated data (Int)). The measurements are accuracy, AUC, FPR, and FNR. Additionally, since the ANN consists of some adjustable parameters, the optimum parameters set, which are the number of input nodes (i), the number of hidden nodes (h), the number of output nodes (o), learning rate (l), and momentum (m), are also reported.

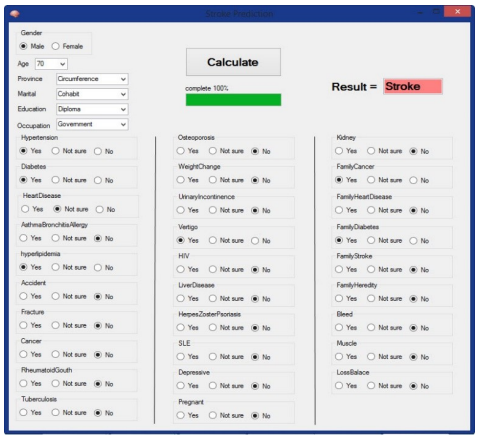

Figure 4: A sample data input for stroke prediction application.

The primary model selection criteria have been set to the measurement of accuracy and AUC because this study emphasizes on the measurement of correctly predicted results. The results in Table 3 are illustrated in a bubble chart shown in Figure 2 which aims to visualize both indicators in a single diagram. Since the ideal model have to satisfy for the high value of both accuracy and AUC, the expected bubble should be near to the top-right corner of the chart while the size of each bubble varies directly with the value of accuracy and AUC. It is shown that there are two best models which can be candidates. The result of the ANN with the integrated data (0.84 accuracies and 0.90 AUC) is very close to the ANN with medical screening data (0.85 accuracies and 0.88 AUC). Although the medical screening dataset provides a slightly higher discrimination power than the integrated data, it is less accurate, based on the whole test set. However, they are much closed. Thus, the secondary criteria are considered.

In the same visualizing fashion, both values are also comparatively illustrated as a bubble chart in Figure 3. The secondary criteria focus on the two types of wrong prediction, the FPR, and the FNR. Ideally, the

FPR can be high for the prevention advantage. On the other hands, the ideal FNR have to be minimized due to the false-diagnostic prevention which should be pessimistic, specifically in the medical domain. Thus, the expected bubble of this comparison should be near to the bottom-right corner of the chart while the size of each bubble varies directly with the FPR but varies inversely with the FNR.

As shown in Figure 3, although the primary measurement of ANN applied with integrated data is close to the medical data, the secondary criteria of these rivals indicate that the ANN with integrated data is superior to another one because of its significant less false negative rate (0.31 and 0.25). This result affirms that the ANN with integrated data set should be awarded as the best stroke prediction model for demographic and medical screening input data due to its capabilities for complication, Therefore, to illustrate the deployment phase, the best ANN model is selected and exemplified with a sample case in the next section.

3.2. Model deployment

Approaching to the practicality, we combine the best features to create a tool and illustrate some input which is simulated as a new patient information shown in Figure 4. For example, the tool can predict that a male patient with age equal seventy years old, living in a suburb, cohabit marital status, hypertension, diabetes, hyperlipidemia, vertigo, and family members appear to have cancer and diabetes, would be predicted as stroke. However, the proposed tool is intended to be applied in ANN model needed for input of all factors. To facilitate end-users, a number of input factors might be reduced if some feature selection methods are applied before model building phase.

4. Conclusions

This research proposed the stroke risk prediction model using three datasets: demographic data, medical screening data and their integration applied by three classification algorithms which are Naïve Bayes, Decision Tree, and Artificial Neural Network (ANN). This research was approved by IRB and data owner to gather and collect outpatients' information from the physical therapy center in Thailand between 2012 and 2015. The demographic data was queried from existing relational database with some filtering criteria while the medical screening data was collected from the paper-based document using a data collecting application. Several versions of medical screening form were integrated into a single data format. The benefit of employing the medical screening data in this research is its user-friendly checklist data form to the patient. After gathering, the imbalance between stroke and non-stroke patients have been found and was solved by the under-sampling method. Then, the data were used in model building by three algorithms and evaluated by the ten-fold cross-validation method. The accuracy and AUC are the primary criteria for model selection while the FP rate and FN rate are the validating criteria in medical field research. The best model from the experiment is the ANN with integrated data by accuracy 0.84, FP rate 0.12, FN rate 0.25 and AUC 0.90. Finally, the deployment is proposed by applied the best model to the simple data input application.

However, since this research developed a stroke risk prediction model based on the data collected from the physical therapy center during 2012 – 2015 which locates near to the capital city, the bias in the data source may be hidden due to the behavior of patients. Additionally, since to medical screening data has been directly input by the patients, the linguistic bias may be attached to the research because some patients may not clearly understand the definition of the medical terms stated in the medical screening form.

To extend and enhance the research to be more realistic and practical, the researcher offers the future works. The first one is to expand the data source to discover a model that representable for the whole population of the country. Another possible future research is that of deeper data exploration and analysis for the benefit of the best resampling method selection. Finally, the feature selection or extraction process may be applied to enhance the model performance.

5. Acknowledgment

This work was supported by the Faculty of Physical Therapy, Mahidol University, Thailand.

References

[1] K. K. Andersen, T. S. Olsen, C. Dehlendorff, L. P. Kammersgaard, Hemorrhagic and ischemic strokes compared, Stroke 40(2009) 2068–72.

[2] World Health Statistics 2015 [Internet], World Health Organization, World Health Organization; 2016 [cited Mar 1, 2018]. Available from: http://www.who.int/gho/publications/world_health_statistics/2015

[3] I. H. Witten, E. Frank, M. A. Hall, C. J. Pal, Data mining: practical machine learning tools and techniques, Amsterdam: Morgan Kaufmann (2017).

[4] G. Piatetsky-Shapiro, W. Frawley, Knowledge discovery in databases, Menlo Park, CA: AAAI Press (1991).

[5] N. Poungvarin, Stroke. 2nd ed. Bangkok: Siriraj hospital (2001).

[6] S. Hanchaiphiboolkul, N. Poungvarin, S. Nidhinandana, N. C. Suwanwela, P. Puthkhao, S. Towanabut, *et al*. Prevalence of stroke and stroke risk factors in Thailand: Thai Epidemiologic Stroke (TES) study, Journal of the Medical Association of Thailand 94(2011) 427–36.

[7] A. K. Arslan, C. Colak, M. E. Sarihan, Different medical data mining approaches based prediction of ischemic stroke, Computer Methods and Programs in Biomedicine 130(2016) 87–92.

[8] L. Amini, R. Azarpazhouh, M. T. Farzadfar, S. A. Mousavi, F. Jazaieri, F. Khorvash, *et al.*, Prediction and control of stroke by data Mining, International Journal of Preventive Medicine 4(2013) s245–s249.

[9] A. Sudha, P. Gayathri, N. Jaisankar, Effective analysis and predictive model of stroke disease using classification methods, International Journal of Computer Applications 43(2012) 26–31.
[10] J. F. Easton, C. R. Stephens, M. Angelova, Risk factors and prediction of very short term versus short/intermediate term post-stroke mortality: A data mining approach, Computers in Biology and Medicine 54(2014) 199–210.
[11] S. Panzarasa, S. Quaglini, L. Sacchi, A. Cavallini, G. Micieli, M. Stefanelli, Data mining techniques for analyzing stroke care processes, Studies in Health Technology and Informatics 2(2010) 939–43.
[12] International Classification of Diseases, 10th Revision (ICD-10) [Internet]. World Health Organization. World Health Organization; 2010 [cited Jun 16, 2018], Available from: http:// www.who.int/ classifications/icd/en/
[13] C. X. Ling, V. S. Sheng, Class imbalance problem, Encyclopedia of Machine Learning and Data Mining (2017) 204–5.
[14] S. Russell, P. Norvig, Artificial intelligence: A modern approach, S.L.: PEARSON (2018).
[15] H. Jeffreys, Scientific inference. Cambridge: Cambridge University Press (2010).
[16] J. R. Quinlan, C4.5 - programs for machine learning. San Mateo, CA: Kaufmann (1992).
[17] S. Kullback, Information theory and statistics. Mineola, N.Y: Dover Publications (1997).
[18] A. Amani, D. Mohammadyani, Artificial neural networks: Applications in nanotechnology, Artificial neural networks - Application. Nov 2011.
[19] L. Franco, J. M. Jerez, J. M. Bravo, Role of function complexity and network size in the generalization ability of feedforward networks, Computational Intelligence and Bioinspired Systems Lecture Notes in Computer Science (2005) 1–8.
[20] R. Eberhart, P. Simpson, R. Dobbins, Computational intelligence PC tools, Boston: AP Professional (1996).
[21] P. A. Devijver, J. Kittler, Pattern recognition: a statistical approach, Taipei: Sung Kang (1982).
[22] G. J. McLachlan,K-A. Do, C. Ambroise, Analyzing microarray gene expression data, Hoboken, NJ: Wiley-Interscience (2004).