Health Risk Analysis Expert System for Family Caregiver of Person with Disabilities using Data Mining Techniques

Ureerat Suksawatchon[1], **Jakkarin Suksawatchon**[2], and **Wannarat Lawang**[3]

ABSTRACT

The nursing care for the family caregiver of the disabled person is an important task for long-term care, since the caring people with disabilities is the difficult and hard task. In this paper, the Health Risk Analysis System or *HRAS* is introduced which is the new expert system for identifying the health risk level in three aspects including mental, physical, and social health aspects, and provides the intervention according to the health risk level of each aspect as well. The *HRAS* is the client-server system. The *HRAS* client proceeds on web-based application to collect health data via online questionnaire and shows the analysis results. The collected health data are transmitted to the server to analysis and to assess the health risk level by using the proposed classifier model named Risk Analysis Classifier or *RAC*. The classification algorithm and rule-based classifier are used to build the *RAC*. The *RAC* is evaluated using k-fold cross validation and the experts with annotated health data and unseen data. The evaluation results showed that Neural Network performs the best performance overall which it achieves the accuracy above 90% in all health data sets. Thus, the Neural Network is the most suitable classifier for this work. In addition, the *HRAS* has been deployed and collected the user experience via the formal survey. These survey results demonstrated that the system provides high accuracy assessment and very utilization in several aspects.

Keywords: health care system, caregiver, classifications

1. INTRODUCTION

There is an increase in adults with a physical disability in developing countries like Thailand. The adults with a physical disability need to find someone to care for them and to support them at home, because paid personal care attendants are one option but it is expensive. In other word, family members become an essential support as "family caregivers". Over 95% of all Thai people with disabilities receive care at home and the majority of caregivers are family members. Therefore, their care is solely dependent on family caregivers whose substantial commitment to long-term care can impact significantly on their health [1]. The caregiver caring is also important for long-term care. Because the caring for a person with disabilities is difficult and hard tasks to provide home care services every day [2]. The effect is that most family caregivers suffer health problems and require support from the people around them.

The research [3] studied the impact of sleep interruptions of the female caregiver and the research [4] studied the impact of caring for a person who has experienced stroke. These researchers found that most family caregivers experienced health problems because of lack of exercise including lack of annual health examination and having sleep trouble. These abusive behaviors required support from the people around them. Especially, caregivers were more inappropriately behave and had health problems than those who did not serve as caregivers of the disabled people. Moreover, the perceived health status was lower than that of the normal population. There are several factors affect the low perception of health status among caregivers in Thailand, including female caregivers to care the mother of the husband, older, low level of education, insufficient income, having a health problem, lack of caring experience, having other roles [5]. Therefore, the assessment of the health risks of the family caregiver is essential. Because the assessment of the factors can indicate the risks of family caregivers in case of health problems and can lead to activities for supporting healthcare [2].

In addition, our team interviewed the expert who has studied in the family care domain since 2013. Our finding from the interview found that typically, the health risk assessment of each caregiver is collected and analyzed by nurses who work in such area and are close to the family caregiver. Each assessment is performed by using the specified questionnaire composed of 6 parts and 143 topics. Therefore, quantitative and quality data are collected concurrently involving a face-to-face interview with family caregivers which spend times to collect the information in approximately 30-45 minutes per person. Besides,

Manuscript received on January 9, 2018 ; revised on June 22, 2018.
Final manuscript received on June 25, 2018.
[1,2] The authors are with Faculty of Informatics, Burapha University, Chonburi, Thailand., E-mail: ureerat@go.buu.ac.th and jakkarin@go.buu.ac.th
[3] The author is with Faculty of Nursing, Burapha University, Chonburi, Thailand., E-mail: lawang@go.buu.ac.th

it spends the time for assessment approximately several hours or several days depend on the experience of the nurses. In addition, each nurse who collected and analyzed the data has experience in different assessment and analyze. The different results will affect the caregiver to get the different treatments.

There are a few researchers developing the risk assessment tools which are close by our work. In recently, the Health-Related Quality of Life (HRQOL) assessment for the caregiver of Alzheimer disease patient [6] was proposed. This tool showed that caregivers who care Alzheimer patients have a poor physical and mental health that yield low quality of life. The HRQOL also assessed the caregiver of autism children [7] and compared the result between caregiver of autism children and general US population. This result showed that the caring of autism children affected to mental, physical health of the caregiver, especially, the caregiver who the mother of autism children. Beside, another example of assessment of the caregiver was proposed in 2008 by Marijean Buhse [8]. The caregiver burden in families of persons with multiple sclerosis was assessed. The finding was that caregiver burden is a multidimensional response to physical, psychological, emotional, social, and financial stressors associated with caregiving experience [8]. Early perception of caregiver burden is very important in determining appropriate interventions. Although there are the variety tools designed for caregiver assessment, however, those tools are not suitable for evaluation the caregiver for the disabled person because of the different culture and context.

Therefore, this research introduces the new expert system named *Health Risk Analysis System* (*HRAS*) for assessing the family caregivers of people with physical disabilities via web-based application. In this system, we propose the Risk Analysis Classifier model called *RAC* by using the classifier technique incorporating with rule-based classifier. The data collected from [1] is used to learn and build the proposed *RAC* model. The proposed model is used to analyze caregiver health risks in three aspects including physical, mental, and social healths. Therefore, the proposed *HRAS* can identify the level of health risk status and an urgent needs for support their health to enable them to maintain their role. In addition, we have chosen the web-based platform for utilization our classifier model. Finally, our work not only helps nurses to easier collect the health but also to reduce analysis and assessment times.

2. LITERATURE REVIEWS

There are various data mining methods which apply to healthcare monitoring. Most of these applications used classifier algorithms such as decision tree, Naïve Bayes, association rules, neural networks and ensemble classifications. For example, Rathore et. al. [9] used ensemble technique to create the predicting model for the survivability of breast cancer patients. This method consists of two steps. First, data preprocessing is applied on SEER (Surveillance of Epidemiology and End Result) data to fill the missing value with the mean value. Second, three classifications are used to build the ensemble classifier model based on voting strategy, including decision tree classifier (DTC), Naïve Bayes, and Classification based on Multiple Association Rules (CMAR). The experimental results were compared with traditional classification. It concluded that the proposed method yielded the highest accuracy at 71.87%, while DTC, Naïve Bayes and CMAR provided the accuracy of 70.00%, 69.00% and 68.20%, respectively.

Besides, Salih et. al. [10] proposed ensemble method based on the Meta classifier voting combining with three based classifiers J48, Random Forest, and Random Tree algorithms. In data processing stage, the input training data were reduced the attributes from 300 to 6 attributes by removing irrelevant and redundant attributes. This work considered the classifier performance using Ensemble Model of Meta Voting Classifiers combining with various single Meta classifiers, Voting combining: J48, LMT, Random Forest, Random Tree, PART (Voting + 5 classifiers), Voting combining: J48, Random Forest, Random Tree (Voting + 3 classifiers), and Voting combining: Random Forest, Random Tree (Voting + 2 classifiers). All proposed methods were compared with various measures of evaluation based on error metrics including ROC curves, Confusion Matrix, Sensitivity, Specificity and the Cost/Benefit measures. They concluded that the ensemble (Voting + 3 classifiers) yielded the highest accuracy .

Srinivas et. al. [11] proposed the intelligent and effective heart attack prediction methods using classification data mining techniques. Firstly, all attributes with missing values are replaced by modes and mean. Then, the discretization of all numerical attribute from the training data set are applied unchanged on the test set. These classification techniques include One Dependency Augmented Naïve Bayes (ODANB), and Naïve Bayes classifier were compared with three data set such as Heart-c, Heart-h and Heart-stalog by using accurate measurement. From experimental results showed that Naïve Bayes provides better results for predicting the heart disease than other methods.

In addition, Songthung and Sripanidkulchai [12] proposed type 2 diabetes mellitus risk prediction by comparing two classifiers including Naïve Bayes and CHAID (Chi-squared Automatic Interaction Detector) Decision tree. The data were gathered from 12 hospitals in Thailand during 2011-2012 with 22,094 records. The coverage and high-risk metrics were used as a measure to compare effectiveness of risk prediction. The results showed that Naïve Bayes classifier obtains a coverage and high-risk which are lower than Decision tree classifier.

Such surveys applied the classification algorithms to produce tools for healthcare applications and can help in assisted healthcare monitoring. Nevertheless, those tools are not suitable for evaluation family caregivers for disabled persons because the surveys are utilized across the target samples in the different culture and context.

3. BACKGROUND KNOWLEDGE

3.1 Classification Techniques

Classification is one of the data mining techniques that widely applied in healthcare applications. The goal of classification is to create a classifier model used to identify the target class for each case in the data. The following techniques are applied in this work.

3.1.1 Decision tree

The goal is to build the classifier model that can predict the value of the target output by learning from the training data. Normally, the decision tree forms a tree-like graph. Each node in the trees specifies a test of some attributes of the instance [13], each branch denoted as the possible values for this attribute, and each leaf node denoted as a class label. In order to select which candidate attributes to be the internal node in the tree, the information theory can be used to measure how well a given attribute separates the training instance. There exist many well-known algorithms, e.g C.4.5, ID3, CART, etc.

3.1.2 Naïve Bayes classifier

Naïve Bayes classifier is simple probabilistic classifiers based on Bayes theorem that suppose the independence of particular feature from other features when given the class. Bayes theorem applies the posterior probability of class (c) when given the attribute or factor (x), $P(c|x)$. Naïve Bayes classifier assume that the effect of the value of an attribute on a given class, is independent of the values of other attributes [13]. In this work, kernel density estimation (KDE) is used for calculating the conditional probability in the Bayes theorem. The KDE is a non-parametric method of estimating the probability density function population [14].

3.1.3 Neural network

In artificial intelligence, neural network is an emulation of a biological neural system of humans in computational system. It builds the model which mimics brain behaviors and be able to learn by the examples or the training data. The basic neural network is consisted of three layers: input layer, hidden layer, and output layer. Each layer have the number of nodes and nodes of input layer are fully connected to the nodes of hidden layer. In the same way, nodes of hidden layer are fully connected to the nodes of output layer. The connections between layers represent weights between nodes. The neural network learning algorithm is to adjust the weight of each neuron that minimizes the average squared error using gradient descent [15].

3.1.4 Support Vector Machines [16]

Support Vector Machines or SVMs are one of the famous machine learning algorithms for regression, classification and other learning tasks such as estimation. LibSVM [16] is a popular open source machine learning library for support vector classification based on C-SVC and also supports multiclass classification. Given a training set of annotated data $(x_i, y_i), i = 1, \ldots, l$ where $x_i \in \Re^n$ and $y \in 1, -1^l$. SVC solves the problem with the following optimization.

$$min_{w,b,\xi} \quad \frac{1}{2}w^T w + C\sum_{i=1}^{l}\xi_i \tag{1}$$

subject to

$$y_i(w^T\phi(x_i + b) \geq 1 - \xi_i \tag{2}$$

$$\xi_i \geq 0, i = 1, \ldots, l,$$

where $\phi(x_i)$ maps x_i into a higher dimensional space and $C > 0$ is the regularization parameter.

3.2 Ensemble Classification

Ensemble classification is the famous technique based on concept of combining classifiers to obtain better predictive performance than using individual classifier alone [10]. Since combining multiple classifiers produce the classifier outputs, so majority voting is one of well-known method used to obtain the true predictive output. The ensemble classifier applied majority voting chooses the class on which all classifiers agree or predicted at least one more than a half of the number of classifiers.

4. HEALTH RISK ANALYSIS SYSTEM

The proposed expert system is illustrated in Fig. 1 called *HRAS* which stands for the Health Risk Analysis System for caregiver of disabled person. The *HRAS* is designed based on the front-end (client) and back-end (server) in order to evaluate the health risk level, and to present the results and appropriate interventions to the user. The front-end of *HRAS* runs on web-based application to present the questionnaire form used to collect related data and transmits the data to the server for processing the assessment. The questionnaire is utilized by nurse who face-to-face interviews the family caregivers. After transmitting the data, the back-end performs preprocessing step for

data cleaning and transforming the categorical data to numeric data. Then the transformed data is sent on a risk analysis modeling which identifies the health risk level. The risk level is stored in a database for further monitoring and then is accessed via the web interface to provide the interventions corresponding the risk level to each user through their account. The interventions can help nurses to support caregiver in the responsible area.

5. *HRAS* WEB INTERFACE

The *HRAS* currently supports only web-based platform with responsive design. The web interface is extraordinarily simple and has three major components including setting permission component, health risk analysis component, and results and report component.

5.1 Setting Permission Component

First time, users are prompted to create a security account on the server by filling out the registration form. The users in the *HRAS* are categorized into two types – the nurse who acts as administration and the nurse who assesses the health risk and supports the caregivers in the responsible area. The primary responsibility of the nurse who acts as the administration is to assign the nurse who will support and monitor the caregivers. Therefore, the web interface utilizes each user with access to the *HRAS* and provides the users with the health risk results from *HRAS*'s database through a personal interface.

5.2 Health Risk Analysis Component

This component is used to assess caregivers' health in three aspects including physical, mental, and social health in the form of the health risk level for each aspect. The assessment is done by the nurse who assesses and supports the caregivers. The nurse obtains the assessment documentation via online questionnaire about demographic and health information, which is consisted of 4 parts in 72 topics. There are only 45 topics related to the important factors and other topics will be further used in the future. This means that spending time to collect the data can be reduced. Because the number of questions is diminished from 143 topics to 72 topics which are selected by the expertise based on research of [1]. After finishing the face-to-face interview, all raw data are kept in the database and sent to analysis (located under "Analysis" Menu) the health risk on the server.

5.3 Results and Report Component

This component is used to visualize the health risk assessment results for the caregiver on the web interface. The *HRAS* can identify the level of health caregiver status into 0-5 levels. The level 0 indicates that health assessment is very good health and the level 5 is urgent need to improve health. The *HRAS* currently supports two health reports. Fig. 2 provides health risk assessment report in four aspects: mental health, physical health, social health and overall heath (from the left hand side to the right hand side). The risk assessment results of each aspect are represented by different emoticon corresponding the risk level and each level is converted to a percentage for easy understanding. When the nurse clicks on emoticon of each aspect, the system will show the intervention designed upon participatory action research (PAR) [1]. Another report is the historical health risk assessments which the nurse can track whether the health of the caregiver can improve, as shown in Fig. 3.

6. HEALTH RISK ANALYSIS PROCESS

After the server obtains the data from the web interface, the *HRAS* back-end performs health risk analysis process to assess the health risk level in three aspects including mental, physical and social aspects. Each step of health risk analysis process is automatically processed in the following.

6.1 Preprocessing

The preprocessing step aims to transform the data obtained from the online questionnaire to the data which can compute and analyze the health risk level. In the online questionnaire, 45 questions are related to the important factors or attributes in each three aspects that are mental health, physical health, and social health. The related factors of each aspect are derived from the research [1] and each factor is obtained from only one question or several questions. The factors related with physical health of the caregiver, are consisted of 11 factors including age, gender, education, relationship with the patient, family economic status, caregiver health problems, care receive-Dependency, caregiving experience, caregiving other role, preparation, ongoing support. The factors that related with mental health of the caregiver, are consisted of 10 factors including age, gender, relationship, family economic status, caregiver health problem, care receive-Dependency, caregiving experience, caregiving hours, preparation, ongoing support. The factors for social health analysis are consisted of 11 factors including age, gender, education, relationship, family economic status, caregiver health problems, care receive-Dependency, caregiving experience, caregiving other role, preparation, ongoing support.

Since all these factors have different types of values, so factor values must be transformed to numeric values. Some factors especially having numeric values, e.g. age, caregiving hours, etc., vary in values. In addition, The values of some factors, e.g care receive-Dependency etc., are obtained from the summation of score calculating all of sub-questions in such factors. Therefore, the values of each factor must be

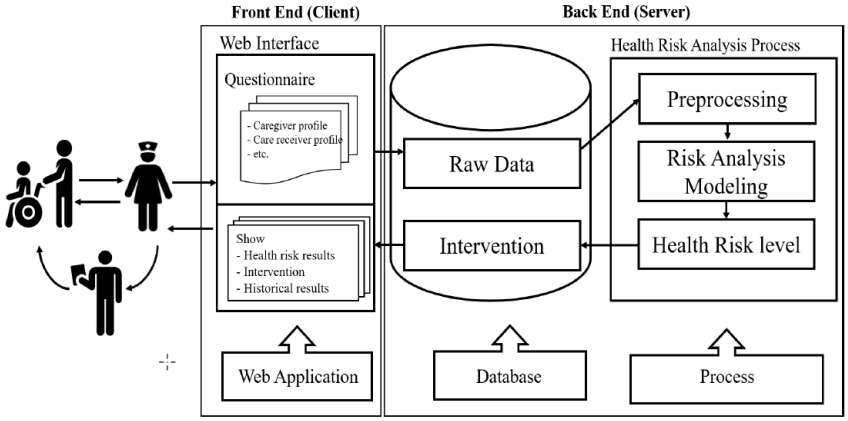


Fig.1: *The Health Risk Analysis System (HRAS) framework.*

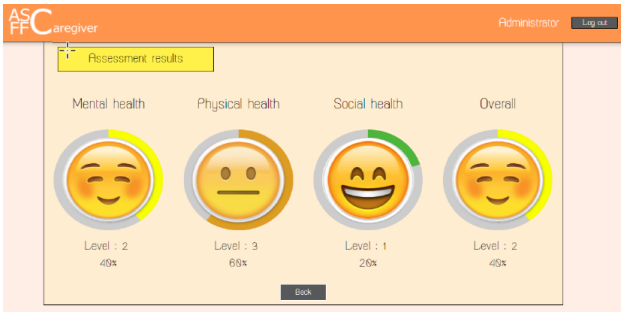


Fig.2: *The assessment result of health risk in four aspects: mental health, physical health, social health and overall heath.*

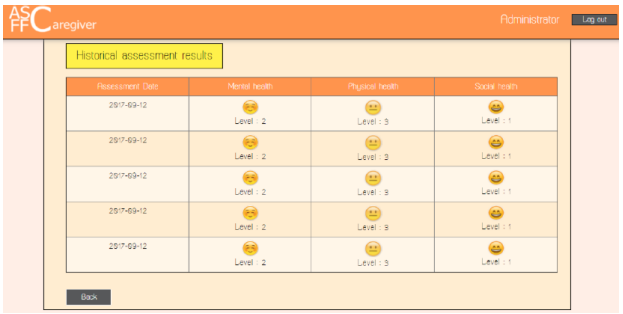


Fig.3: *The historical assessment result.*

converted to the specific value. Table 1 shows the example of some factors related to all health risks. Since some factors used to assess influence the risk assessment of each aspect in a different way, for example, the age factor more effects in physical health than mental health. Thus, the weight notion is used for all factors. All weights are obtained from [1]. To finding the weight, they used T-test, Z-test, and Pearson correlation for finding the correlation between factors and poor health including the depth interview with 15 experts. Table 2 illustrates the example of data transformation of one caregiver in mental health data. After that, data of a caregiver acquired from online questionnaire are transformed to the numeric data for convenience calculation as depicted in Fig. 4.

6.2 Risk Analysis Modeling

This step is used to identify the health risk level in three aspects including mental health, physical health, and social health by using *Risk Analysis Classifier* or *RAC* which is the mixed classifier model. In order to create the *RAC*, classification technique and rule-based method are applied to build the proposed classifier model. Fig. 5 illustrates the process of building the *RAC*.

Caregiver	F1	F2	F4	...	F7	F8	...	F14
C1	0.4	1	1.5	...	3	1.5	...	0
C2	0.6	1	0		0	1.5		3
C3	0	0	0		3	0		3
C4	1	1	3		3	2		0
C5	0.8	0	0		0	2		1.5

Fig.4: *The example of transformed data of five users in mental health data. The values are obtained from the total scores as shown in Table 2.*

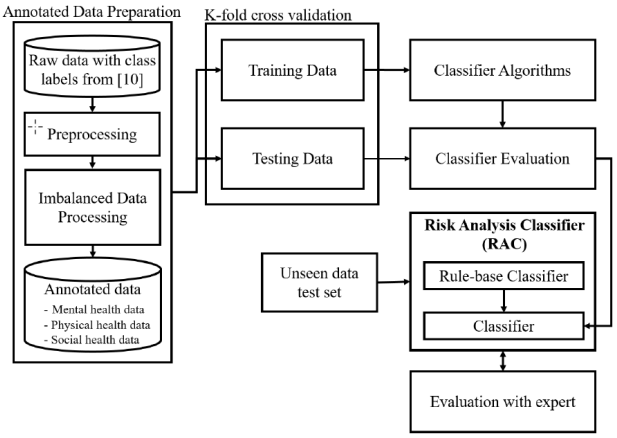


Fig.5: *The overview processes of building Risk Analysis Classifier or RAC.*

Table 1: *Some factors related with all health and weight of each factor*

Factors	Convert to	Weight of factors		
		Mental	Physical	Social
Age (F1)				
< 30 years.	0			
30-39 years.	0.2			
40-49 years.	0.4	1	3	-
50-59 years.	0.6			
60-69 years.	0.8			
⩾ 70 years.	1			
Caregiver health problem (F7)				
Healthy	0	3	3	3
Poor health	1			
Care receive-Dependency (F8)				
Score dependency is 17 - 20	0			
Score dependency is 13 - 16	0.25			
Score dependency is 9 - 12	0.5	2	2	2
Score dependency is 5 - 8	0.75			
Score dependency is 0 - 4	1			
Preparation (F14)				
Score is 11 - 12	0			
Score is 8 - 10	0.5	3	2	3
Score is 6 - 7	1			

Table 2: *Data transformation in preprocessing of one caregiver with mental health data*

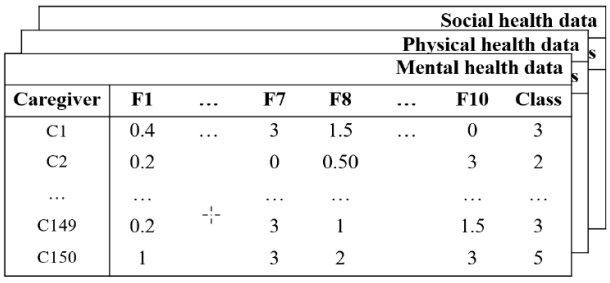
Factors	Values of factor	Convert to	Weight	Total Score
Age (F1)	42	0.4	1	0.4
Caregiver health problem (F7)	Poor health	1	3	3
Care receive-Dependency (F8)	8	0.75	2	1.5
Preparation (F14)	12	0	3	0

Table 3: *The number of instances of each dataset*

Class	Mental health data		Physical health data		Social health data	
	Number of instances	Percent of instances	Number of instances	Percent of instances	Number of instances	Percent of instances
2	9	6%	11	7.33%	28	19.05%
3	47	31.33%	54	36%	67	45.58%
4	68	45.33%	64	42.67%	48	32.65%
5	26	17.33%	21	14%	4	2.72%

6.2.1 Annotated Data Preparation

The *RAC* is built using training data from 150 caregivers with annotated risk levels obtained from the research [1]. Since each caregiver will be assessed in three aspects that mentioned before, so the training data are generated into three data sets according to three aspects. The example of some annotated data is shown in Fig. 6. Normally, the level scale identifying the health risk level should be determined 1 to 5 scale. Since, the training data obtained from the research [1] provide the annotated data only 2 to 5 scale, so this causes the *RAC* model combining with two techniques that are classification and rule-based methods. In addition, the training data are questionnaire data as well, therefore, the data transformation is applied using the **Preprocessing** step in Section 6.1.



Caregiver	F1	…	F7	F8	…	F10	Class
C1	0.4	…	3	1.5	…	0	3
C2	0.2		0	0.50		3	2
…	…		…	…		…	…
C149	0.2		3	1		1.5	3
C150	1		3	2		3	5

Fig.6: *The examples of annotated data of mental, physical and social aspects.*

From our previous work [2], we found that the size of one or more classes (risk levels) are much more greater than the other classes as shown in Table 3. Thus *RAC* model tends to majority classes while the data from the minority classes may be incorrectly classified. It means that *RAC* model encounters with a crucial problem called *class imbalance*. Therefore, we have the assumption that if we can handle the class imbalance problem, we can obtain highly accurate *RAC* model. In this work, we applied *SCUT* [17] algorithm to cope with class imbalance problem. The reason of selecting this method is that the SCUT method is the hybrid sampling technique used to balance the size of training instances in a multi-class imbalance data. Furthermore, the SCUT method combines both undersampling and oversampling tech-

niques. For all classes having the number of instances less than the average m, oversampling is proceeded by using SMOTE [18]. The percentage of sampling is computed which the number of instances in the class after oversampling is equal to the average m. Otherwise, for all classes having the number of instances greater than the average m, undersampling is performed such that the number of instances is equal to the average m. For undersampling, the Expectation Maximization (EM) algorithm [19] is used to cluster the instances within each class. Then, for each cluster within the class, instances are uniformly selected which the number of instances for all cluster is equal the average m. The algorithm of SCUT method is shown in **Algorithm SCUT** [17].

Since SCUT algorithm applies random selection in the undersampling process, so the selected instances of each cluster may be difference from each run. To obtain the best result, we perform 5 runs to create various datasets. Therefore, there are 18 annotated health datasets used for experiments in this work that are 3 health datasets used before performing SCUT method and 5 datasets for each three aspect used after performing SCUT method (totally 15 datasets).

Algorithm SCUT
Input: Dataset D with n classes.
Output: Dataset $D^{'}$ with all classes having m instances, where m is the average number of instances of all classes.

Split D into $D_1, D_2, \ldots, D_i, \ldots, D_n$ where D_i is a single class and $i = 1, 2, \ldots, n$.
Compute m. /*Undersampling*/
For each D_i where number of instances $> m$
  Cluster D_i using EM algorithm.
  For each cluster C_j, $j = 1, \ldots, k$
    Randomly select instances from C_j.
    Add selected instances to $C_j^{'}$.
  End For
  $C = \phi$
  For $j = 1, \ldots, k$
    $C = C \cup C_j^{'}$.
  End For
  $D_i^{'} = C$
End For
/*Oversampling*/
For each D_i where number of instances $< m$
  Apply SMOTE on D_i to get $D_i^{'}$.
End For

For each D_i where number of instances $= m$
  $D_i^{'} = D_i$.
End For
$D^{'} = \phi$
For $i = 1, \ldots, n$
  $D^{'} = D^{'} \cup D_i^{'}$.
End For

Return $D^{'}$.

6.2.2 Building Risk Analysis Classifier (RAC)

The RAC model is a hybrid model by combining with two techniques which are rule-based classifier and classification methods. In this work, the ruled-based is utilized for identifying the risk levels 0 and 1 by sum scores of all factors. If the total score equals to 0 then the health risk level is 0. If the total score is greater than 0, but less than 4 then the health risk level is 1. These rules are obtained from the expert. The other levels will identify by the suitable classifier model provided by the experiments. Since, there are many classification methods used to build classifier model. In this research, we perform the experiment with different classifier algorithms and different parameters. Four classification algorithms are selected including Decision tree (C4.5), Naïve Bayes using kernel density estimation, Feed-forward neural network (Back propagation) with sigmoid function, Support Vector Machine or SVM using LibSVM [16] and Ensemble classifiers with voting. For decision tree, C4.5 is selected because it is improved from ID3 by adding pruning process for cutting the unimportant branches. Besides, three classifier algorithm with the high accuracy are chosen to perform the Ensemble classifier with voting. For the utilization, the classification algorithm with the highest performance is selected as the RAC model.

Table 4: *Parameters setting of each classification algorithm*

Models	**Model types**	**Parameters**
Decision Tree	C4.5	confidence : [0.25, 0.5] scaled up by 0.25. minimum size for splitting : [4, 20] scaled up by 4.
Naïve Bayes	Kernel	minimum bandwidth : [0.1, 1] scaled up by 0.1. number of kernels : [10, 100] scaled up by 10.
Neural Network	Back propagation	training cycles : [500, 1000] scaled up by 100. leaning rate : [0.1, 1] scaled up by 0.1. momentum : [0.1, 1] scaled up by 0.1.
LibSVM	C-SVC	degree : [1, 10] scaled up by 1. epsilon : [0.001, 0.01] scaled up by 0.001.
Ensemble with Voting	Kernel Back propagation C-SVC	using the same previous parameters of three models

6.2.3 Evaluation

In this research, the model evaluation was performed on two cases including classifier evaluation and evaluation the RAC with the experts. The objec-

Table 5: *Accuracies of each classifier algorithm*

Health Data	Classification techniques				
	Decision Tree	**Naïve Bayes (Kernel)**	**Neural Network**	**LibSVM**	**Ensemble with voting**
Mental					
without SCUT	72.00 ± 6.99%	76.14 ± 8.56%	92.00 ± 5.81%	82.11 ± 6.06%	87.32% ± 8.42%
with SCUT	74.30% ± 5.88%	76.94% ± 10.28%	94.71% ± 6.52%	87.50% ± 5.84%	88.81% ± 5.38%
Physical					
without SCUT	74.78 ± 6.06%	78.59 ± 7.85%	90.67 ± 6.11%	84.00 ± 3.27%	84.07% ± 8.27%
with SCUT	76.32% ± 5.88%	82.27% ± 6.73%	91.17% ± 7.39%	93.42% ± 5.06%	90.81% ± 4.00%
Social					
without SCUT	72.79 ± 6.60%	76.14 ± 6.95%	93.88 ± 3.33%	78.33 ± 6.39%	85.71% ± 5.09%
with SCUT	77.67% ± 8.84%	76.26% ± 6.94%	95.36% ± 2.55%	84.92% ± 10.62%	88.24% ± 6.79%

tive of classifier evaluation is to choose the most suitable classification algorithm among four algorithms. All these algorithms were compared by four measurements of classification performance metrics such as accuracy, precision, recall, and F-measure [20]. The parameters of each classification algorithm were set and vary in value as shown in Table 4. For Ensemble classifier, this work selected the three classifiers that give the best accuracy. In addition, the notion of k-fold cross validation was used to evaluate the classifier models. In this work, we also performed the experiments by varying the parameter k for each data set and each algorithm. After the suitable classification algorithm was obtained, we also evaluated the *RAC* model using unseen data of three health data sets, and then examined the classified results by the experts.

Table 6: *Confusion matrix for Neural Network with mental health dataset*

		Actual			
		Class 2	**Class 3**	**Class 4**	**Class 5**
Predicted	**Class 2**	38	0	0	0
	Class 3	0	37	2	0
	Class 4	0	1	32	1
	Class 5	0	0	4	37

Table 7: *Confusion matrix for Neural Network with physical health dataset*

		Actual			
		Class 2	**Class 3**	**Class 4**	**Class 5**
Predicted	**Class 2**	38	2	0	0
	Class 3	0	31	2	0
	Class 4	0	5	30	0
	Class 5	0	0	6	38

Table 8: *Confusion matrix for Neural Network with social health dataset*

		Actual			
		Class 2	**Class 3**	**Class 4**	**Class 5**
Predicted	**Class 2**	38	1	0	0
	Class 3	0	36	0	0
	Class 4	0	0	33	0
	Class 5	0	1	5	38

7. RESULTS AND DISCUSSIONS

The summary results for our experiments are shown in Tables 5. This table presents the accuracies from previous work [2] before combining with the SCUT algorithm, and the accuracies after combining with the SCUT algorithm to deal with class imbalance problem. Table 5 shows the health risk prediction experiments performed in three health datasets and displays the best predictive accuracies associated with each of health datasets, for each of five the classification algorithms. Since the results are obtained by k-fold cross validation, so the results are presented in the average accuracy with variance of all experiments.

Table 5 demonstrates that each classifier algorithm combined with the SCUT algorithm gives the accuracies higher than the accuracies obtained from the classifier algorithms without the SCUT algorithm. From this table, we can conclude that the SCUT algorithm can handle the class imbalance problem and gives the high accuracy values. In addition, Table 5 demonstrates that Neural Network can achieve the high accuracies overall rank for all three health datasets. While Ensemble Classifier with voting ranks second for mental health and social health datasets. On the other hand, LibSVM ranks second only physical health data. Besides, Neural Network does outperform overall rank which it can achieve accuracy above 90% in all health data sets. Because Table 5 indicates that Neural Network can deal with the nonlinear data and overlapping data better than other algorithms. Therefore, Neural Network is the most suitable in this case and it is chosen to be the classifier in *RAC* model. More detailed results are presented in Tables 6 - 8, which display the confusion matrices for Neural Network combined with the SCUT algorithm by selecting the best result from all experiments.

Besides, Table 9 - Table 11 show the other performances including precision, recall, and F-measure metrics, respectively. Each of performance metric was computed from the confusion matrix that gave the highest accuracy from all experiments for each classifier algorithm. From these tables, we can see that the most results are substantially improved when the classifier algorithms are incorporated with the SCUT algorithm. Finally, we also evaluated the *RAC* model using the unlabeled data set which consisted

Table 9: *Precision metric for each model with health datasets*

Health Data	Classification techniques				
	Decision Tree	**Naïve Bayes (Kernel)**	**Neural Network**	**LibSVM**	**Ensemble with voting**
Mental					
without SCUT	65.20%	**79.67%**	93.24%	63.69%	86.10%
with SCUT	**74.31%**	73.35%	**94.81%**	**87.72%**	**89.17%**
Physical					
without SCUT	70.55%	**84.09%**	92.95%	65.07%	87.60%
with SCUT	**75.94%**	82.05%	**94.81%**	**93.78%**	**90.63%**
Social					
without SCUT	67.04%	59.03%	70.13%	64.55%	65.18%
with SCUT	**78.07%**	**75.26%**	**95.95%**	**85.25%**	**88.03%**

Table 10: *Recall metric for each model with health datasets*

Health Data	Classification techniques				
	Decision Tree	**Naïve Bayes (Kernel)**	**Neural Network**	**LibSVM**	**Ensemble with voting**
Mental					
without SCUT	67.11%	72.81%	87.60%	64.06%	77.77%
with SCUT	**74.35%**	**75.00%**	**94.74%**	**87.50%**	**88.82%**
Physical					
without SCUT	70.09%	71.98%	78.60%	66.58%	74.10%
with SCUT	**76.32%**	**82.24%**	**90.13%**	**93.42%**	**90.79%**
Social					
without SCUT	63.00%	55.20%	72.84%	53.05%	65.54%
with SCUT	**77.63%**	**76.32%**	**95.40%**	**84.87%**	**88.16%**

Table 11: *F-measure metric for each model with health datasets*

Health Data	Classification techniques				
	Decision Tree	**Naïve Bayes (Kernel)**	**Neural Network**	**LibSVM**	**Ensemble with voting**
Mental					
without SCUT	65.69%	75.69%	89.43%	63.02%	80.60%
with SCUT	**73.25%**	**73.67%**	**94.67%**	**87.29%**	**88.57%**
Physical					
without SCUT	69.81%	76.39%	80.43%	65.37%	78.12%
with SCUT	**75.80%**	**81.92%**	**89.91%**	**93.30%**	**90.60%**
Social					
without SCUT	63.77%	56.10%	71.43%	52.73%	65.30%
with SCUT	**77.81%**	**75.35%**	**95.41%**	**84.32%**	**87.77%**

of 30 samples, and then examined the classified results with two experts as shown the process in Fig. 5. The experts have examined that are "agree" and "disagree" in the predicted risk level for each sample provided by the proposed system. The examinations with experts gave accuracy in 80% for mental and physical data sets and up to 90% for social data set. In addition, both experts also gave the positive opinions especially that they confirmed that "the proposed system can help the nurses in analysis and reducing the assessment time". This means those results are acceptable to the experts.

8. CONCLUSIONS

In this paper, the Health Risk Analysis System or *HRAS* is introduced for assessment the caregiver's health in term of health risk levels in mental health, physical health and social health. The objective of this system is to analyze and assess the health risk level of the family caregivers, to provide the health reports and intervention to support the family caregivers, and to track the health improvement. To analysis and assess the health risk level, the new Risk Analysis Classifier or *RAC* is proposed which combining with the classification technique and rule-based classifier. The proposed classifier *RAC* has tested using k-fold cross validation and evaluated with two experts who work in the family caregivers domain as well. The experiments performed with annotated data and unseen data sets. The evaluation results demonstrated that Neural Network does outperform for overall data sets which it achieves all metrics above 90%. Moreover, the *HRAS* has been utilized and collected the user experience via formal survey questionnaire. The results showed that the *HRAS* can provide the nurses who are the focus group, analyzing the health risk level with an accuracy assessment and reducing the assessment time.

ACKNOWLEDGMENT

First of all, this research work was supported by faculty of Informatics, Burapha University, Thailand. Secondly, we would like to thank to our graduate student, Miss Chalaruk Kritsanaphuti, who helped us in completing this research. Finally, we would also like to thank Mr. Hemmarat Wachirahatthaphong, who

helped us to set up the back-end environment.

References

[1] W. Lawang, D.E. Horey, and J. Blackford, "Family caregivers of adults with acquired physical disability: Thai case-control study," *Int J Nurs Pract*, vol. 21, no. 1, pp. 70-77, 2015.

[2] C. Kritsanaphuti, U. Suksawatchon, W. Lawang, and J. Suksawatchon, "Health risk analysis system for family caregiver of disabled person," in *The 2nd International Conference on Information Technology (InCIT2017)*, Nov 2017, pp. 120-125.

[3] K. Tsukasaki, T. Kido, K. Makimoto, R. Naganuma, M. Ohno, and K. Sunaga, "The impact of sleep interruptions on vital measurements and chronic fatigue of female caregivers providing home care in Japan," *Nurs Health Sci*, vol. 8, no. 1, pp. 2-9, Mar 2006.

[4] K. Salter, L. Zettler, N. Foley, and R. Teasell, "Impact of caring for individuals with stroke on perceived physical health of informal caregivers," *Disabil Rehabil*, vol. 32, no. 4, pp. 273-281, 2010.

[5] S.R. Beach, R. Schulz, G.M. Williamson, L.S. Miller, M.F. Weiner, and C.E. Lance, "Risk factors for potentially harmful informal caregiver behavior," *J Am Geriatr Soc*, vol. 53, no. 2, pp. 255-261, Feb 2005.

[6] M. I. Andreakou, A. A. Papadopoulos, D. B. Panagiotakos, and D. Niakas, "Assessment of Health-Related Quality of Life for caregivers of Alzheimers disease patients," *International Journal of Alzheimers Disease*, http://doi.org/10.1155/2016/9213968.

[7] R. Khanna, S. S. Madhavan, M. J. Smith, J. H. Patrick, C. Tworek, and B. Becker-Cottrill, "Assessment of health-related quality of life among primary caregivers of children with autism spectrum disorders," *J Autism Dev Disord*, vol. 41, no. 9, pp. 1214-1227, Sep. 2011.

[8] M. Buhse, "Assessment of caregiver burden in families of persons with multiple sclerosis," *Journal of Neuroscience Nursing*, vol. 40, issue 1, pp. 25-31, Feb 2008.

[9] N. Rathore, Divya, and S. Agarwal, "Predicting the survivability of breast cancer patients using ensemble approach," in *2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, Feb 2014, pp. 459-464.

[10] A. S. M. Salih and A. Abraham, "Novel ensemble decision support and health care monitoring system," *Journal of Network and Innovative Computing*, vol. 2, no. 2014, pp. 041-051, 2014.

[11] K. Srinivas, B. K. Rani, and A. Govrdhan, "Applications of data mining techniques in healthcare and prediction of heart attacks," *International Journal on Computer Science and Engineering (IJCSE)*, vol. 2, no. 02, pp. 250-255, 2010.

[12] P. Songthung and K. Sripanidkulchai, "Improving type 2 diabetes mellitus risk prediction using classification," in *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, July 2016, pp. 1-6.

[13] T. M. Mitchell, "Machine learning and data mining," *Commun. ACM*, vol. 42, no. 11, pp. 30-36, Nov. 1999. [Online]. Available: http://doi.acm.org/10.1145/319382.319388

[14] Y. Murakami and K. Mizuguchi, "Applying the nave bayes classifier with kernel density estimation to the prediction of proteinprotein interaction sites," *Bioinformatics*, vol. 26, no. 15, pp. 1841-1848, 2010.

[15] G. P. Zhang, "Neural networks for classification: a survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 30, no. 4, pp. 451-462, Nov 2000.

[16] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, Article 27 (May 2011), pp. 27:1-27:27 DOI=http://dx.doi.org/10.1145/1961189.1961199

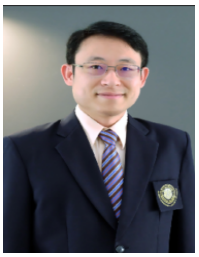
[17] A. Agrawal, H. L. Viktor and E. Paquet, "SCUT: Multi-class imbalanced data classification using SMOTE and cluster-based undersampling," in *the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, Lisbon, 2015, pp. 226-234.

[18] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer, "SMOTE: synthetic minority oversampling technique," *J. Artif. Int. Res*, vol. 16, no. 1, pp. 321-357, June 2002.

[19] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," Paper presented at the Royal Statistical Society at a meeting organized by the Research Section, December 8, 1976.

[20] D.M.W. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37-63, 2011.

Ureerat Suksawatchon received the B.Sc. degree in Mathematics from Kasetsart University, Thailand in 1997. She received M.Sc. and Ph.D. degrees in Computer Science from Chulalongkorn University, Thailand in 2001 and 2008, respectively. She is a lecturer at Faculty of Informatics, Burapha University, Thailand. Currently, she is an Assistant Professor in Computer Science, Faculty of Informatics, Burapha University. Her research areas are recommendation systems, digital image processing, intelligent system, pattern recognition and data analytics.



Jakkarin Suksawatchon received the B.Sc. degree in Computer Science from Burapha University, Thailand in 2007, and M.Sc. degree in Information Technology, King Mongkut's Institute of Technology Ladkrabang, Thailand, in 2001. He received Ph.D. degree in Computer Science, Chulalongkorn University, Thailand in 2007. He is currently a lecturer at Faculty of Informatics, Burapha University, Thailand. He is an Assistant Professor in Computer Science, Faculty of Informatics, Burapha University. His current research works concern recommendation systems, digital image processing, data analytics, and pattern recognition especially activity recognition.

Wannarat Lawang received the Bachelor degree of Nursing Science in 1995 and Master degree of Nursing Science in Community Health Nursing in 1999. She received Doctor of Philosophy in Health Sciences from La Trobe University, Australia in 2013. During 1995 - 1997, she worked the Cancer and Radiological Nursing Unit, Siriraj Hospital, Thailand. Currently, she is a lecturer and an Assistant Professor at Division of Community Nursing, Faculty of Nursing, Burapha University. Her research interest are health care systems and population including caregivers, disabled persons.