Data Mining Approaches in Personal Loan Approval

Kuannabha Pimcharee[1] and Olarik Surinta[2,*]

[1]Department of Information Technology, Faculty of Informatics, Mahasarakham University, Mahasarakham, Thailand

[2]Multi-agent Intelligent Simulation Laboratory (MISL), Department of Information Technology, Faculty of Informatics, Mahasarakham University, Thailand

olarik.s@msu.ac.th[*] (corresponding author)

Abstract. *The approval of a bank's credit for an individual loan requires the fulfillment of several requirements, such as bank credit policy, loan amount, the purpose of the loan, and repayment ability. However, every type of credit is subject to the risk of non-repayment and non-performing loans, which affect the liquidity of the bank's operation. This research studied the application of data mining techniques to identify key factors for the loan decisions of a bank. The main objective was to compare the data mining process of personal loan approval process with and without feature selection techniques. For the experiments, the first step was to create the data mining models using three methods, including support vector machine (SVM), multi-layer perceptron (MLP), and decision tree. The results showed that the SVM method outperformed other data mining methods. Second, we experimented with feature selection techniques consisting of Chi-square and information gain. The Chi-square considered the ten factors, while information gain selected the best three factors. The experimental results showed that the Chi-square and information gain combined with the MLP method obtained an accuracy rate of 90.40% and 91.70%, respectively. Therefore, this research concluded that the SVM classifier without combining the feature selection method is the best method to use in personal credit evaluation.*

Received by	17 June 2021
Revised by	13 July 2021
Accepted by	17 July 2021

Keywords:

Support Vector Machine, Multi-Layer Perceptron, Decision Tree, Feature Selection, Chi-square, Information Gain, Personal Loan Approval

1. Introduction

Non-repayment is a risk for all types of loans, resulting in non-performing loans (NPLs). According to the Bank of Thailand (BOT), all banks are required to set up reserve money for doubtful debts in compliance with the BOT's policy. The policy has an impact on a bank's liquidity. Therefore, good credit management reduces the risk of outstanding debt. Good credit management includes effective analysis of loan approval, tracking of debt collection and future customers' repayment performance. This issue covers the analysis process for debt compromises. Good credit management requires accounting data as an instrument for financial statement analysis, business operating results, liquidity, previous business performance, present & future business plans [1]. This information is used for credit consideration and approval.

The People's Bank Project (Government policy credit) [2] is a government established project intended to increase financial loan opportunities for retail customers. The project's goal is to respond to both government policy and the vision of the government's savings bank. As a result, different criteria for loan approval are required for each type of loan. The bank uses different policies and guidelines for each customer. The main factors are 1) policy for each type of credit. For example, some banks may require a borrower to have a job for more than one year or to own a business. 2) The loan amount is appropriate and in accordance with the purpose of the loan as stated in the application form. For example, the loan money will be used to purchase a home. 3) Characteristics and repayment ability of the borrower. Normally, financial institutions use the 5Cs criteria as instruments. The 5Cs are as follows. - 1) Character, which is the borrower's reputation or the borrower's information. 2) Capacity, which is a borrower's ability to repay a loan at an allocated time. 3) Capital, which lenders consider the capital of the borrower, e.g. cash, mortgages, etc. 4) Collateral, which can be people or objects, which helps lenders ensure that they will be able to recover their money from the collateral. 5) Conditions are the external factors which affect the borrowers' income. All banks collect these five factors for credit evaluation before approval.

In response to the information presented above, we are concerned about the factors affecting credit approval. Therefore, we collected data from the bank's borrowing applicants from 2017 – 2018, which totals 1,000 items. All of the transactions were approved transactions and unapproved transactions. Then, this data was introduced to the data mining process to create a model and classify the data by Support Vector Machine (SVM), Multi-Layer Perceptron (MLP) and Decision tree. After that, we used feature selection methods which included Chi-square and information gain to identify the factors affecting the credit approval. The selected factors were used for data modeling by SVM, MLP, and decision tree to find the suitable process for personal loan credit evaluation.

Paper Outline: Section 2 will describe related research on credit evaluation. Section 3 discusses the People's Bank Project, which was initiated by the Government Saving Bank. The project allows people to access funds and borrow money from informal loans. Section 4 discusses data mining techniques in the experiments. Section 5 describes the characteristics of the collected data which has been used to create a model. Section 6 describes the result of the experiment and Section 7 discusses the summary of the experiment.

2. Related Work

Data Mining is one of the effective instruments for the evaluation of credit approval. The Data Mining techniques which are commonly used for modeling are J48, Bayesian network, Naive bayes, Logistic regression Neural networks, and Radial basis function (RBF). All of these techniques are used for credit evaluation and loan risk assessment in banks, credit risk assessment for commercial banks and setting scores for credit scoring [3-7].

Ince & Aktan [5] compared four data mining techniques, including discriminant analysis, logistic regression, neural network, and classification and regression tree (CART) for credit scoring. The 1,260 records were tested by those instruments. The details are as follows. The 890 records are training sets and the 370 records are a test set. The experiments show that CART achieved the most predicted accuracy rate, at 65.58%. The neural network method achieved the lowest predicted accuracy rate, at 61.52%. Moreover, Wah & Ibrahim [6] used data mining techniques including logistic regression (LR), classification and regression tree (CART) and neural network to create a model for classification of credit card approval data. The experiment shows that the neural network method achieved the highest rate of classification at 76.46% while the LR method and CART method achieved the rate of classification at 74.56% and 73.66% respectively.

Bekhet & Eletter [7] developed a credit risk assessment model for commercial banks which used 492 cases from Jordanian commercial banks as tested data. Data were collected between 2006 and 2011 and included 292 creditworthy cases and 200 unworthy cases. There were 12 attributes; age, gender, total income, company's type, guarantor, loan amount, loan purpose, period with the current employee, duration of credit, nationality, interest rate, and debt payment ratio. For the experiment, the data was tested by logistic regression (LR) and the radial basis function (RBF) scoring model. The result showed that LR is a better method than RBF for the overall accuracy rate. In contrast, the RBF method is more accurate in cases of identifying the customer's grace period. After that, Hamid & Ahmed [8] developed a model for credit evaluation of loan risk in banks. The research used 1,000 records from the bank to render a model. The bank's data consisted of 8 attributes, including credit history, purpose, gender, credit amount, age, housing, job, and class. Data was divided into two parts for data modeling. The data for the training set accounted for 80% of the total, while the data for the test set accounted for 20%. The research tested the data with three algorithms, including J48, Bayesian network, and the Naive Bayes. The result showed that J48 got the highest accuracy rate at 78.38%.

For feature selection before modeling data, Suksomboon & Jongkasikit [9], used evolutional selection, which is the process of reducing the dimensions of the data (Attribute selection). The process helps the corporative to analyze the repayment ability of the cooperative members. There were a total of 19 attributes in the analysis. After the analysis, there were 12 important attributes which contributed most to the evaluation. These were history of stock accumulation (12 months), number of people under care, loan history (number of times), loan period, monthly income, marital status, cooperative member period, amount of recent loans, repayment history (12 months), amount of stocks held, cash deposit history (12 months), and monthly expenses. Each attribute had a value ranging from 0.892 to 0.304 and was sorted from most important to least important. The decision tree, Naive Bayes, and K-nearest neighbors were used to test the data. The decision tree had the highest accuracy rate, at 95.96%, according to the results.

Sitthisan & Seresangtakul [10] introduces multiple regression analysis to analyze the relationship between each attribute for data modeling. The created model was used for prediction of education officers' repayment ability. Only eight of the 14 attributes had an impact on repayment ability. Those attributes were the number of children, number of repayment (period), age, amount of loan, monthly income, remaining debt, insurance description, and gender. The correlation analysis used significance levels of 0.01 and 0.05, which means the confidence intervals were 99% and 95%, respectively. After that, we then tested the data by using a neural network, Bayesian belief network, and decision tree by cross-validation. The K value has been set to K=10. The result shows that the three methods have an accuracy rate at 98.7%.

3. Bank's Process for Credit Approval

The criteria of financial institutions for credit approval depend on the lenders' policies and criteria. In general, the main factors which are used for consideration are lenders' policies, purpose of the loan, the characteristics and the repayment ability of the borrower [1]. Most financial institutions use the 5Cs principle. The 5Cs are as follows.

1. Character is the characteristic of the borrower. It refers to the repayment history of the borrower, which reports repayment ability and debt management.

2. Capacity is the repayment ability of the borrower compare to the borrower's income. It also refers to the stability of the borrower's income.
3. Capital is fund, assets, and cash which the borrower has own. It is used as capital of the loan.
4. Collateral refers to an asset or person that a lender accepts as security for a loan.
5. Conditions refers to external factors which affects the borrowers' income

In general, credit reviewers calculate the borrowers' data as follows;

Net income = Salary + Extra income (1)

Total cost = Debt + Payment (2)

Amount of debt: The People's Bank Loan Project is a loan that uses a fixed interest (Flat Rate), which the borrowers pay fixed rate interest from total principal throughout the contract period.

Interest = Principal x Interest rate% x Number of periods

Payment = (Principal + Interest)/(Number of periods) (3)

The equation (3) divides data into 2 groups, which are:

1. A group that has the ability to repay (approved groups).
2. A group that does not have the ability to repay (unapproved groups).

The process above is a basic process for a bank to distinguish the ability to make a repayment. This data is used to create actual labels for the data used in the experiment.

4. Bank's Process for Credit Approval

Data mining is the process of extracting large amounts of data, discovering patterns and correlations using machine learning methods. Data mining processes include preprocessing, data transformation, data selection, data analysis for pattern detection, evaluation, and interpretation [11]. This study looks at feature selection techniques and classification algorithms in the following ways.

4.1 Feature Selection Techniques

4.1.1 Chi-Square

Chi-Square is a non-parametric method which is used for testing the independence hypothesis between categorical variables of 2 variables or above [13]. The Chi-square test tells researchers whether a variable is independent to label or not. The Chi-square equation is shown as (4).

$$x^2 = \frac{\sum_{i=1}^{k}(o_i - E_i)^2}{E_i} \quad (4)$$

where O is observed values, E is expected values, and i is the i^th position in the contingency table.

4.1.2 Information Gain

The value obtained from information gain is the difference between x, which is a target variable and A, which is an independent variable [14]. Information gain decreases the entropy of the target variable x by learning from the status of the independent variable A.

Information gain preferentially selects attributes with a high value of differences.

4.2 Classification Algorithms

This research uses three classification algorithm techniques for data modeling in the case of banks' credit approval evaluation. These are support vector machine (SVM), multi-layer perceptron (MLP), and decision tree.

4.2.1 Support Vector Machine (SVM)

In 1998, Vapnik [15] proposed a support vector machine (SVM) for a data classification model. SVM is a two-class classification [16]. The SVM is based on statistical learning [17]. Recently, it has been commonly used in several fields related to image classification such as pedestrian detection [18], [19], remote sensing image classification [20], etc.

SVM performs classification by finding the hyperplane that maximizes the margin between the two classes (positive and negative). ผิดพลาด! ไม่พบแหล่งการอ้างอิง shows the optimal hyperplane (*H*), which classifies positive and negative. The range is the maximum distance of the margin between the hyperplane and support plane *H1* and *H2*) and is called the maximum margin or decision boundary. As a result, the data *x* is known as "support," and it is located on the support plane (support vector). The decision function is given by:

$$f(x) = sign(w^T x + b) \quad (5)$$

SVM is intended to solve linear problems, which are presented as equations:

$$H: w^T x + b = 0 \quad (6)$$

$$H1: w^T x + b = -1$$

$$H2: w^T x + b = 1$$

where *w* is weight vector to the hyperplane, *x* is input vectors, and *b* is a bias value. Then $\frac{2}{\|w\|}$ is the size of the margin. The linear kernel function is defined as follows:

$$k(x_i, x_j) = x_i^T x_j \tag{7}$$

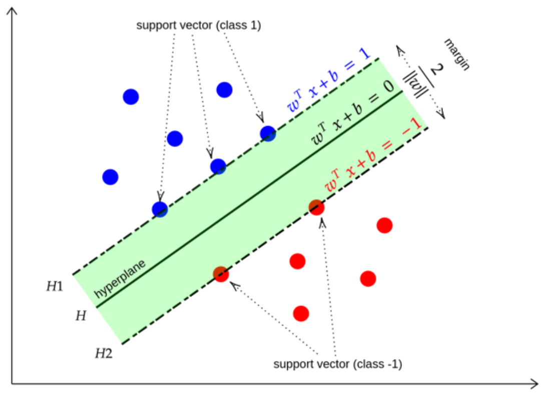


Fig. 1 Illustration of the optimal hyperplane and support vectors

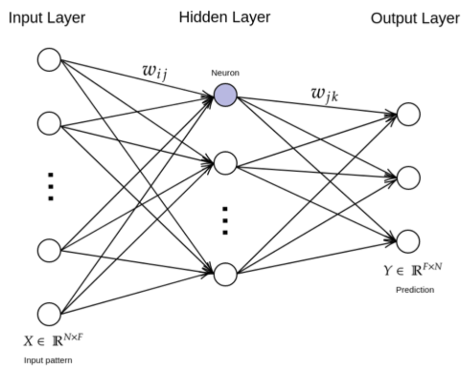
4.2.2 Multi-layer Perceptron (MLP)

Multi-layer Perceptron (MLP) is supervised learning method and popular neural network architecture for classification work [21]. The MLP method imitates the structure and behavior of neurons in the human brain. The layers (input layer, hidden layer, and output layer) of each neuron are linked to one another. The topology is fully connected (see Figure 2a). Weight will be used to calculate each neuron or unit. The weight sum is calculated by the activation function (see Figure 2b). The activation function maps the weighted inputs to the output of each neuron and uses them as the input of neurons in the next layer. When error values occur during data training, the network weight will be updated during the calculation process. The error value is calculated from the predicted value of the network and the actual value of the network. Therefore, the network learns from the dataset and the data label.

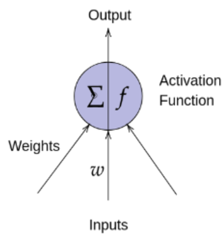
4.2.3 Decision Tree

J. Ross Quinlan invented the decision tree, known as the ID3 [22], in 1986. The method was designed to imitate a tree structure which consists of root nodes, branches, and leaf nodes. Following that, C4.5 was created, which included a greedy algorithm for classification work.

Creating a tree requires probability distribution calculation of all possible events (Entropy) (P), assigned $P = (p_1 = \frac{|c_1|}{|T|}, p_2 = \frac{|c_2|}{|T|}, ..., p_k = \frac{|c_k|}{|T|},)$, T is a set of training examples that belongs to the class $c_i, i = 1, ..., k$ In general, entropy values are computed as follows:



(a)



(b)

Fig. 2 Illustration of the multi-layer perceptron: (a) the network consists of an input layer, a hidden layer, and an output layer and (b) simple computational units that have a weighted input signal and produce an output signal using an activation function

$$I(P) = -\sum_{i=1}^{k} p_i \cdot \log(p_i) \tag{8}$$

The following equation is used to calculate the information gain value:

$$Info(X,T) = -\sum_{i=1}^{k} \frac{|T_i|}{|T|} \cdot Info(T_i) \tag{9}$$

$$Gain(X,T) = Info(T) - Info(X,T) \tag{10}$$

where T_i is the partition of T included by the value of X.

4.3 Evaluation Metrics

Evaluation metrics are an assessment of the category's predictive performance of personal loan approval. This study evaluates the method's performance in terms of accuracy, precision and recall [23], as shown in the equations below:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{11}$$

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

The value of the result from the classification is:

TP is the amount of data that correctly identifies customers with repayment ability (True Positive).

TN is the amount of data that correctly identifies customers who are unable to repay their debts (True Negative).

FP is the amount of data that incorrectly identifies customers with the repayment ability (False Positive).

FN is the amount of data that incorrectly identifies customers who are unable to repay their debts (False Negative).

5. Personal Loan Data Set

We collected 1,000 loan application data from 2017–2018 for this study. There were 500 loan applications approved by the banks, while another 500 loan applications were unapproved. As shown in Table 1, there were 14 relevant attributes in bank loan applications.

Item	Attributes	Data Type
1	Age	integer
2	Gender	binominal
3	Occupation	polynominal
4	Marital status	polynominal
5	Salary	integer
6	Extra income	integer
7	Cost	integer
8	Net income	integer
9	Cost of living	integer
10	Debt	integer
11	Payment	integer
12	Total cost	integer
13	Loan amount	integer
14	Number of periods	integer

Table 1 A total of 14 attributes are considered by the bank when making a loan decision

6. Experimental Results

The experiment used 1,000 instances which are were classified by the K-Fold Cross-Validation method [24]. The experiment assigned K=10 in a training set and test set. The results of the experiment are as follows:

6.1 Evaluation Performance on Machine Learning Techniques

The 14 features were used with machine learning methods, including SVM, MLP, and decision tree methods, in order to find the best method which had the highest accuracy rate. The assessment results were used in credit evaluation. Model performance was evaluated by standard deviation: SD, precision, and recall value.

Optimization was performed in this experiment to find the most suitable parameters for the personal loan dataset. The parameter that was used with SVM was the Linear Kernel. The suitable value was 59.59. Subsequently, we used MLP to generate the number of nodes in the hidden layer. The optimal number of nodes was 21 nodes. The criterion in the decision tree method is defined as the gain ratio, and the maximum depth value was set to 10. The results are displayed in Table 2.

Methods	Evaluation Metrics		
	Accuracy	Precision	Recall
MLP	83.99± 4.08	83.07 ± 5.98	85.98 ± 7.04
SVM	**97.12± 2.83**	**96.42 ± 3.97**	**97.99 ± 1.97**
Decision tree	60.06± 4.36	59.14 ± 3.50	64.30±11.57

Table 2 The comparison of 3 machine learning techniques by the cross-validation method of 14 features

The experiment reports SVM using a linear kernel with parameter C optimization. The optimal C value in this experiment is was 59.59. The highest accuracy rate among the three methods was the SVM method at 97.12%, which was 14% more than the MLP method. The decision tree method is not suitable for personal loan evaluation because the accuracy rate was only 60.06%.

6.2 Evaluation Performance on Feature Selection

The 14 attributes were put through the feature selection process in order to select the suitable features using the Chi-Square method and information gain. The results of the feature selection process are as follows:

6.2.1 Performance using Chi-Square Method

When all 14 attributes are calculated by Chi-Square, the weight of each feature is shown in ผิดพลาด! ไม่พบแหล่งการอ้างอิง. The weight value from ผิดพลาด! ไม่พบแหล่งการอ้างอิง is considered. The analysis employs the reliability value (*df*). The reliability value is assigned at *df*= 0.01. The weight value which is considered must be greater than 27.7. Therefore, this experiment uses the attribute numbers 1-10 respectively for data modeling by SVM, MLP, and decision tree.

Item	Features	Weighted Value
1	Total cost	131.679
2	Debt	128.632
3	Net income	61.284
4	Number of periods	51.351
5	Cost	50.511
6	Salary	49.884
7	Occupation	46.822
8	Payment	42.466
9	Loan amount	36.923
10	Loan amount	33.188
11	Cost of living	23.154
12	Marital status	22.785
13	Extra income	4.946
14	Gender	1.210

Table 3 Weighted value of each feature which calculated by Chi-square method

Methods	Evaluation Metrics		
	Accuracy	Precision	Recall
MLP	**90.40± 2.67**	**90.40± 5.89**	**90.40 ± 2.79**
SVM	89.40± 6.64	90.87± 4.66	87.60 ± 3.89
Decision tree	62.10± 3.73	64.44± 6.47	54.00± 12.45

Table 4 Attributes selection by Chi-Square and personal loan credit evaluation by machine learning techniques

According to Table 4, attribute selection using Chi-Square yields 14 attributes for data modeling using MLP, SVM, and decision tree. The experiment found that MLP and SVM achieved the highest accuracy rate for personal loan credit evaluation at 90.40% and 89.40% respectively. The decision tree, on the other hand, is an ineffective method for evaluating personal loans due to its low accuracy rate. The accuracy rate of the decision tree is 62.10%.

6.2.2 Performance using Information Gain Method

The information gain method selected 3 suitable attributes which were net income, debt, and total cost. These three attributes were used for data modeling by MLP, SVM, and decision tree.

Methods	Evaluation Metrics		
	Accuracy	Precision	Recall
MLP	**91.70± 3.59**	**90.49± 5.09**	**93.20 ± 6.04**
SVM	91.20± 1.69	88.72± 2.00	94.40 ± 1.40
Decision tree	51.80± 2.10	51.22± 0.00	75.80± 28.77

Table 5 Attribute selection by information gain and personal loan credit evaluation by machine learning techniques

ผิดพลาด! ไม่พบแหล่งการอ้างอิง shows that the information gain method chose three attributes and then applied them to data modeling. The result shows that the MLP method achieved the highest accuracy rate for credit evaluation at 91.70%, while the SVM method achieved an accuracy rate of 91.20%. Therefore, the experiment concludes that both MLP and SVM can be used for credit evaluation.

7. Conclusion

Data mining techniques are used in this paper for credit evaluation in the bank's credit approval process. Credit approval data of 1000 applicants was tested. The data was classified into two groups; approved loan applications and unapproved loan applications. The experiment used the cross-validation method, assigned K=10 to find the accuracy value. The support vector machine (SVM), multi-layer perceptron (MLP), and decision tree method were all used in data modeling for data classification. Furthermore, we compared the effectiveness of the Chi-Square and information gain methods by using data features from both methods to build a model and classify it using SVM, MLP, and a decision tree. The results showed that the feature selection method is not suitable to be used for credit approval evaluation because the accuracy rate is 91% while using all 14 features gets the higher accuracy rate at 97%.

In the future, we intend to experiment with feature selection methods such as factor analysis, LightGBM, and XGBoost to find suitable features [25-27]. Ensemble learning methods [28] can be used for bringing different machine learning models into the credit approval process. This may result in a higher accuracy rate for the credit approval process.

References

[1] "Bank of Thailand Conditions financial institutions use to consider credit." [Online]. Available: https://www.1213.or.th/th/serviceunderbot/loans/Pages/beforeloan.aspx. [Accessed April 2021].

[2] "The People Bank Project", Loans of government savings bank." [Online]. Available: https://www.gsb.or.th/gsb_govs/bank_people/. [Accessed May 2021].

[3] R. Wang, "AHP-entropy method credit risk assessment based on Python." *Proc – Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS)*, pp. 17-12, 2021, doi: 10.1109/ACCTCS52002.2021.00011.

[4] Y. Deng, Y. Wei, and Y. Li, "Credit risk evaluation based on data mining and integrate feature selection," *Proc – International Conference on Signal Processing, Communications and Computing (ICPCC)*, pp. 1-4, 2020. doi: 10.1109/ICSPCC50002.2020.9259483.

[5] H. Ince and B. Aktan, "A comparison of data mining techniques for credit scoring in banking: A managerial perspective," *J. Bus. Econ. Manag.*, vol. 10, no. 3, pp. 233–240, 2009, doi: 10.3846/1611-1699.2009.10.233-240.

[6] Y. B. Wah and I. R. Ibrahim, "Using data mining predictive models to classify credit card applicants," *Proc - 6th International Conference on Advanced Information Management and Service (IMS)*, pp. 394–398, 2010.

[7] H. A. Bekhet and S. F. K. Eletter, "Credit risk assessment model for Jordanian commercial banks: Neural scoring approach," *Rev. Dev. Financ.*, vol. 4, no. 1, pp. 20–28, 2014, doi: 10.1016/j.rdf.2014.03.002.

[8] A. J. Hamid and T. M. Ahmed, "Developing prediction model of loan risk in banks using data mining," *Mach. Learn. Appl. An Int. J.*, vol. 3, no. 1, pp. 1–9, 2016, doi: 10.5121/mlaij.2016.3101.

[9] P. Suksomboon and N. N. Jongkasikit. Effects of models that affect their ability to repay loans of member in credit union cooperative. *Industrial Technology Lampang Rajabhat University Journal*, vol. 10, no. 1, pp. 101–110, 2017.

[10] Y. Sitthisan and P. Seresangtakul. "The model of predicting settlement ability: A case study in office of the welfare promotion commission for teachers and educational personnel, Udonthani Province, Thailand," *Journal of Chemical Information and Modeling*, vol. 53 no. 9, pp. 1689–1699, 2018. doi: 10.1017/CBO9781107415324.004

[11] S. Vilailuck, V. Jaroenpuntaruk and D. Wichadakul. "Utilizing data mining techniques to forecast student academic achievement of Kasetsart University Laboratory School Kamphaeng Saen Campus Educational Research and Development Center," *Veridian E-Journal, Science and Technology Silpakorn University*, vol. 2, no. 2, pp. 1–17, 2015.

[12] N. Chanamarn. "Learning attributes analysis by feature selection and prediction," *Sakon Nakhon RajaBhat University Journal*, vol. 6, No. 12, pp. 31–46, 2014.

[13] S. Bahassine, A. Madani, M. Al-Sarem, and M. Kissi, "Feature selection using an improved Chi-square for Arabic text classification," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 32, no. 2, pp. 225–231, 2020, doi: 10.1016/j.jksuci.2018.05.010.

[14] S. Lei, "A feature selection method based on information gain and genetic algorithm," *Proc. - International Conference on Computer Science and Electronics Engineering (ICCSEE)*, pp. 355–358, 2012, doi: 10.1109/ICCSEE.2012.97.

[15] C. Vapnik. "Model complexity control and statistical learning theory," *Nat. Comput.*, vol. 1, no. 1, pp. 109–133, 2002, doi: 10.1023/A:1015007927558.

[16] G. Cheng and X. Tong, "Fuzzy clustering multiple kernel support vector machine," *Proc - International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)*, pp. 7–12, 2018, doi: 10.1109/ICWAPR.2018.8521307.

[17] H. Elaidi, Y. Elhaddar, Z. Benabbou, and H. Abbar, "An idea of a clustering algorithm using support vector machines based on binary decision tree," *Proc - International Conference on Intelligent Systems and Computer Vision (ISCV)*, pp. 1–5, 2018, doi: 10.1109/ISACV.2018.8354024.

[18] M. J. Flores Calero, M. Aldas, J. Lazaro, A. Gardel, N. Onofa, and B. Quinga, "Pedestrian detection under partial occlusion by using logic inference, HOG and SVM," *IEEE Lat. Am. Trans.*, vol. 17, no. 9, pp. 1552–1559, 2019. doi: 10.1109/TLA.2019.8931190.

[19] B. Meus, T. Kryjak, and M. Gorgon, "Embedded vision system for pedestrian detection based on HOG+SVM and use of motion information implemented in Zynq heterogeneous device," *Proc - Signal Processing Algorithms, Architectures, Arrangements and Applications (SPA)*, pp. 406–411, 2017. doi: 10.23919/SPA.2017.8166901.

[20] P. Liu, K. K. R. Choo, L. Wang, and F. Huang, "SVM or deep learning? A comparative study on remote sensing image classification," *Soft Comput.*, vol. 21, no. 23, pp. 7053–7065, 2017. doi: 10.1007/s00500-016-2247-2.

[21] J.P.L. Pizzaia, I.R. Salcides, G.M.d. Almeida, R. Contarato, and R.D. Almeida, "Arabica coffee samples classification using a multilayer perceptron neural network," *Proc - IEEE International Conference on Industry Applications (INDUSCON)*, pp. 80-84, 2018. doi: 10.1109/INDUSCON.2018.8627271.

[22] J.R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol.1, pp.81-106, 1986, doi: 10.1007/BF00116251.

[23] M. Junker, R. Hoch, and A. Dengel, "On the evaluation of document analysis components by recall, precision, and accuracy," *Proc - Fifth International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1-4, 1999, doi: 10.1109/ICDAR.1999.791887.

[24] P. Tamilarasi and R. U. Rani, "Diagonis of crime rate against women using k-fold cross validation through machine learning," *Proc – Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 2034-1028, 2020. doi: 10.1109/ICCMC48092.2020.ICCMC-000193.

[25] Z. Zhang and Y. Zhang, "Evaluation of inter-provincial medical and health service quality and influencing factors based on factor analysis," *Proc – International Conference on Public Management and Intelligent Society (PMIS)*, pp. 54-57, 2021. doi: 10.1109/PMIS52742.2021.00019.

[26] J. Bao, "Multi-features based Arrhythmia diagnosis algorithm using Xgboost," *Proc – International Conference on Computing and Data Science (CDS)*, pp. 454-457, 2020. doi: 10.1109/CDS49703.2020.00095.

[27] F. Ye, J. Wang, Z. Li, Z. Hihan and C. Yang, "Jane street stock prediction model based on LightGBM," *Proc – 6th International Conference on Intelligent Computing and Signal Processing (ICSP)*, pp. 385-388, 2021. doi: 10.1109/ICSP51882.2021.9408851.

[28] S. Noppitak and O. Surinta, "Ensemble convolutional neural network architectures for land use classification in economic crops aerial images," *ICIC Express Lett.*, vol. 15, No. 6, pp. 531-543. doi: 10.24507/icicle.15.06.531.

Biographies

Kuannabha Pimcharee received her bachelor degree in Computer Science from Mahasarakham University, Thailand. She is currently working at Government Saving Bank. Her research interests include data mining and feature selection.

Olarik Surinta is currently a lecturer at Department of Information Technology, faculty of Informatics, Mahasarakham University, Thailand. He received master degree in Information Technology from King Mongkut's Institute of Technology North Bangkok, Thailand. He received his Ph.D. in Artificial Intelligence from University of Groningen, The Netherlands. His research interests are deep learning, machine learning, artificial intelligence, computer vision, and handwritten recognition.